Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Тюменский государственный нефтегазовый университет»

Ш. Ф. Фарахутдинов, А. С. Бушуев

ОБРАБОТКА И АНАЛИЗ ДАННЫХ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ В ПАКЕТЕ SPSS 17.0
КУРС ЛЕКЦИЙ

Учебное пособие
для студентов очной и заочной форм обучения
гуманитарных направлений

Тюмень
ТюмГНГУ
2011

УДК 303.1
ББК С6
Ф 24

Рецензенты:
доктор социологических наук, профессор В. В. Гаврилюк
доктор социологических наук, профессор О. М. Барбаков

Фарахутдинов, Ш. Ф.
Ф 24 Обработка и анализ данных социологических исследований в пакете SPSS 17.0. Курс лекций : учебное пособие / Ш. Ф. Фарахутдинов, А. С. Бушуев. – Тюмень : ТюмГНГУ, 2011. – 220 с.
ISBN 978-5-9961-0414-7

В учебном пособии содержатся сведения по обработке и анализу данных, полученных в результате социологических исследований. Особое внимание уделяется характеристике сущности каждого метода, его эвристическим возможностям и описанию необходимых процедур для его реализации в пакете SPSS 17.0.

Адресованная студентам, профессиональная сфера деятельности которых в будущем может быть связана с социологическими исследованиями, книга может быть полезна для магистрантов, аспирантов и специалистов, работающих над диссертационными исследованиями и не имеющих опыта обработки и анализа данных исследований, а также всем желающим самостоятельно освоить пакет SPSS.

УДК 303.1
ББК С6

ISBN 978-5-9961-0414-7

ОГЛАВЛЕНИЕ

Введение

За последние несколько десятилетий поток окружающей нас информации сильно увеличился. Если раньше знания и навыки, обретенные человеком в процессе социализации и получения образования, оставались актуальными на протяжении длительного периода времени, то сейчас ситуация совершенно иная. Для того чтобы оставаться полноценным членом общества как в бытовой, так и профессиональной сферах, человек должен уметь эффективно работать с большим количеством информации, постоянно осваивать новые знания и навыки.

Не случайно в настоящее время в России осуществляется переход системы образования на стандарты третьего поколения, ключевым элементом которых является компетентностный подход. Освоение компетенций здесь происходит в процессе изучения как отдельных учебных дисциплин, циклов, модулей, так и дидактических единиц, которые интегрируются в общепрофессиональные и специальные дисциплины. Подчеркивается связь рассматриваемого предмета с различными знаниями, умениями, навыками. В этой связи актуализируется древнее высказывание о том, что обучающийся – это не сосуд, который необходимо заполнить, но факел, который нужно зажечь. Настоящий курс лекций, по задумке авторов, призван выполнить именно эту функции.

Учитывая, что статистическая обработка данных эмпирических исследований, проведенных в любой сфере науки – очень важный, но в то же время сложный этап получения нового знания, этому направлению должно уделяться особое внимание. До недавнего времени эта отрасль оставалась прерогативой специалистов-математиков, однако современные реалии требуют владения подобными навыками обработки и анализа данных специалистов других сфер.

Сказанное особо актуально для специалистов-социологов, которые в силу «гуманитарности» своего образования с трудом воспринимают различные математические формулы и статистические показатели. Тем не менее владение этим инструментом является обязательной компетенцией, которая позволяет социологу быть полноценным специалистом.

В настоящее время существует много учебников, пособий, справочной литературы, посвященных обработке и анализу данных социологических исследований. Огромное количество информации по этим вопросам имеется в сети Интернет. Однако, как показывает опыт авторов, разобраться студентам в этом разнообразии бывает крайне сложно. Часть этой информации перенасыщена формулами, математической терминологией, другая грешит отсутствием наглядности, конкретных примеров, а ведь это очень важно для

полноценного понимания рассматриваемого вопроса. Большинство учебников универсальны и рассчитаны на специалистов самых различных отраслей: медиков, экономистов, психологов и др.

Особенность данного издания состоит в том, что оно составлялось исключительно для тех, кто планирует работать с социологическими исследованиями. Соответственно и материал, и примеры были подобраны таким образом, чтобы студенты испытывали как можно меньше затруднений при освоении содержания. В предлагаемом курсе лекций, как правило, не даются детальные математические выкладки, так как это усложнило бы текст и сделало бы более трудным его восприятие студентами-гуманитариями. Основной упор авторами сделан на характеристике сущности каждого метода, его эвристических возможностях и описании необходимых процедур для его реализации в SPSS. Кроме того, книга изобилует скриншотами, которые позволяют без особого труда разобраться в осуществляемых статистических процедурах.

Стоит отметить, что в издании дается информация об источнике эмпирических данных – социологических архивах. Читатель, желающий изучить SPSS, используя данный курс лекций, сможет без труда найти подходящие для выполнения самостоятельных работ первичные данные. Параллельно с тренировкой практических навыков работы в SPSS, работая с архивными базами, у обучающегося есть возможность осуществлять реальные исследования методом вторичного анализа данных социологических исследований. Опыт преподавания авторами дисциплины «Системы обработки данных в социологии» показывает, что результаты таких мини-исследований студентов отлично дополняют курсовые и выпускные квалификационные проекты, а также могут быть самостоятельными работами, публикуемыми в студенческих научных сборниках. Таким образом появляются конкретные навыки и достигается выработка компетенций.

Возможным недостатком данного курса лекций является то, что здесь не раскрыты все имеющиеся возможности программы SPSS. Однако авторы и не преследовали этой цели. Выбор тем, которые нашли свое отражение в лекциях, был обусловлен реальной практикой авторов в этой области. Различные исследования, в которых довелось принять участие авторам, а также стажировки в крупнейших центрах изучения общественного мнения страны показали, что на практике социолог использует довольно ограниченное количество методов. Так, на официальном сайте ВЦИОМ[1] в разделе «Методы исследований, используемые центром» упоминаются следующие: дескриптивный статистический анализ данных, многомерный анализ данных (факторный, кластерный анализ, корреспонденс-анализ, многомерное

---

1 http://wciom.ru/

шкалирование, множественная регрессия), регрессионно-корреляционный анализ, статистическое тестирование гипотез, моделирование, типологизация. Практически весь перечисленный набор, а также работа с данными (подготовка к анализу, вычисление переменных, перегруппировка, категоризация и т.д.), рассмотрены в рамках настоящего курса лекций.

Дополнительную информацию по теме обучающиеся могут найти в специальной литературе. Для этого в конце каждой лекции приведен список рекомендуемой литературы. Также некоторую необходимую информацию можно извлечь из встроенной в программе справки. Для закрепления полученных знаний в конце лекций также приводится список вопросов и контрольных заданий. Самостоятельная работа с ними, безусловно, будет способствовать закреплению знаний и превращению полученных умений в навыки.

Авторы выражают благодарность организаторам и всем преподавателям Центра социологического и политологического образования Института социологии РАН, у которых им довелось учиться весной 2008 года на интенсивных специализированных курсах «Возможности программы SPSS для анализа социологической информации», в частности: Г.Н. Воронину, М.Ф. Чернышу, А.В. Чурикову, С.Е. Кухтерину, И.Ф. Девятко, Л.А. Хахулиной, а также Л.П. Ипатовой.

Лекция 1. Архивы данных социологических исследований

Проблема создания общедоступных архивов социологических данных обрела особую актуальность в конце 40-х гг. XX в. В первую очередь это было связано с огромным и поистине бесценным материалом, накопленным в ходе проведения возрастающего числа социологических исследований, а также пониманием того, что большие деньги, вложенные в получение этого материала, не должны пропасть. В то же время со стороны общества чувствовались возрастающая потребность в доступе к подобной информации, желание знать и анализировать живые фактографические данные, полученные в ходе социологических опросов. Пионерами в области архивирования выступили Институт Гэллапа и вновь организованный центр Ропера, ставший хранителем и распространителем результатов исследований, проведенных Институтом Гэллапа.

В последующие три десятилетия система архивов активно развивалась в США и странах Европы. К настоящему моменту архивы социологических данных, открытые для научного и экспертного сообщества, существуют практически во всех демократических обществах. Они объединены в международные союзы и организации и своей целью ставят создание единого информационного пространства, позволяющего пользователю получить релевантную информацию независимо от места ее физического хранения.

Первые попытки создания электронных архивов в Советском Союзе были предприняты в Институте социологии. В 1985 г. для накопления данных эмпирических социологических исследований и организации их хранения в виде, пригодном для многократного использования, был организован банк данных социологических исследований (БДСИ). Над его созданием работал целый ряд сотрудников Института социологии АН СССР, в том числе В. Т. Андреенков, А. В. Жаворонков, М. С. Косолапов, А. О. Крыштановский, О. М. Маслова, В. Н. Шипиловидр. В 1987 г. банк приобрел статус всесоюзного. В нем аккумулировали информацию социологических исследований, проводимых не только сотрудниками Института социологии АН СССР, но и другими социологическими центрами. В настоящее время банк формально утратил статус всесоюзного, но сохранил эмпирические данные широкого спектра исследований, проводимых в течение длительного периода практически во всех республиках бывшего СССР.

Значительные перемены в условиях, обеспечивающих развитие открытых электронных архивов социологических данных, произошли в конце 90-х гг. XX в. К этому времени уже существовало достаточно социологических агентств, работающих по международным стандартам и получающих достоверную и важную информацию о жизни общества. Более того, поя-

вились вспомогательные средства — компьютеры, сети, программное обеспечение, облегчающие доступ к этой информации. И, наконец, возникли необходимые социальные условия, делающие возможным открытый доступ к информации о жизни общества. На фоне этих перемен в сентябре 2000 г. был создан Единый архив социологических данных (ЕАСД).

Основная цель проекта — обеспечить свободный доступ научного и экспертного сообщества к результатам социологических опросов. Решение этой задачи важно не только для развития социальных наук в России: свободный доступ к достоверным данным — основа развития гражданского общества и демократии. Эта возможность одинаково важна и для ученых, и для студентов, и для средств массовой информации, и для исследовательских организаций. Жизнь в демократическом, открытом обществе невозможна без доступа к достоверной информации, отвечающей на самые актуальные вопросы общественной жизни, позволяющей исследователю понять и сделать понятными для широкой публики важнейшие процессы социальной трансформации. Проект был начат по инициативе АНО «Левада-центр», к нему присоединились ведущие исследовательские организации: Институт социологии РАН, Фонд «Общественное мнение», Институт комплексных социологических исследований, РОМИР, Институт экономики и организации промышленного производства СО РАН, депонировавшие в архив результаты своих исследований.

С 2002 г. Единый архив существовал и работал как специализированная программа в рамках Независимого института социальной политики. В 2010 г. Единый архив социологических данных был объединен с базой данной по экономике государственного университета Высшей школы экономики (ГУ-ВШЭ). Идея создания баз данных по российской экономике была предложена Евгением Григорьевичем Ясиным в 1999 г. В следующем году был учрежден Институт информационного развития, занимавшийся, в том числе, поддержкой и сопровождением баз данных по экономике. В 2001 г. экономические базы данных стали открытым информационным ресурсом, доступ к которому осуществлялся через Интернет, в 2002 г. появилась англоязычная версия системы.

Решение о создании единого научно-исследовательского подразделения было принято в декабре 2009 г. С этого момента начинается новый этап развития архива, который теперь носит название **Единый архив экономических и социологических данных** (ЕАЭСД). Но основные принципы работы — обеспечение сохранности эмпирических данных и организация свободного доступа — остались прежними. Остановимся подробнее на деятельности архива и возможностях его использования в исследовательских и образовательных целях.

Коллекция архива. К настоящему моменту (на конец 2010 г.) в коллекциях Архива накоплено более 750 социологических исследований и более 100 временных рядов основных показателей российской экономики. Это

превышает пилотный этап Архива более чем в 10 раз. Постоянно расширяется и круг депозиторов: к первоначальной шестерке отцов-основателей присоединились и многие другие (табл. 1.1). В настоящее время Единый архив начал работу по сбору результатов исследований получателями грантов благотворительных фондов.

Таблица 1.1

Основные депозитарии Единого архива экономических и социологических данных

	Аналитический центр Юрия Левады (Левада-центр)
ВЦИОМ	ВЦИОМ (Всероссийский центр изучения общественного мнения)
GfK	ГфК «Русь»
CR RC Caucasus Research Resource Centers	Закавказский Ресурсный Исследовательский Центр (CRRC)
	Институт философии РАН
	ИСРАН (Институт социологии Российской академии наук)
ИСЭПН	ИСЭПН РАН (Институт социально-экономических проблем народонаселения Российской академии наук)
	ИЭОПП СО РАН (Институт экономики и организации промышленного производства Сибирское отделение Российской академии наук)
COMCON	КОМКОН
10 лет	Московский центр Карнеги
РОМИР	РОМИР (Российское общественное мнение и исследование рынка)
	Фонд ИНДЕМ
	ФОМ (Фонд «Общественное мнение»)
	Центр «Стратегия» СПб

География пользователей архива. Успешно развивается взаимодействие с пользователями архива. Если в 2003 г., когда начиналась отработка критериев оптимального взаимодействия с пользователями, было заключено 67 соглашений о передаче данных, то в 2006 г. их уже около 200. Единый архив уделяет большое внимание расширению круга пользователей – постоянно осуществляется поиск новых форм работы, организовываются новые сервисы. Возможность работать с данными в режиме on-line особенно важна для региональных пользователей. Это подтверждает география заявок: они поступают практически из всех регионов России. Примерно 10% заявок приходит от зарубежных пользователей (рис. 1.1).



Рис. 1.1. География запросов в Единый архив экономических и социологических данных

Особое внимание Архив уделяет сотрудничеству с высшими учебными заведениями, готовящими специалистов по социально-экономическим и политическим дисциплинам. ЕАЭСД стремится к тому, чтобы материалы архива использовались в учебном процессе в качестве эмпирической основы. Это необходимо для того, чтобы использование материалов Единого архива в учебном процессе и для написания различного рода работ, от студенческих курсовых до диссертаций, стало нормой.

С 2004 г. ведется сотрудничество с экономическим факультетом Московского государственного университета. В ходе серии семинаров были продемонстрированы возможности архива и достигнута договоренность о том, что данные Архива будут использоваться в ходе учебного процесса и для самостоятельной работы студентов. Большая работа ведется с региональными университетами. Так, в Омском государственном университете данные Единого архива не только используются в учебном процессе, но и стали основой конкурса студенческих работ. Активно используются данные Единого архива в учебном процессе Новосибирского государственного университета. В архив постоянно обращаются представители различных региональных вузов.

Структура социологической информации содержащейся в Архиве. Информация по социологической тематике, хранящаяся в архиве, доступна для свободного пользования[2]. Все данные на сайте архива четко структурированы и содержат следующие разделы: «Тематические исследования», «Повторяющиеся исследования», «Объединенные исследования», «Компаративные исследования», «Исследования ГУ-ВШЭ».

В раздел «Тематические исследования» вошли опросы, посвященные изучению конкретных тем. Респонденты отвечали на вопросы о бедности и богатстве, безработице и занятости, государственных социальных программах, критериях достижения успеха, причинах различия в доходах, целях политики, удовлетворенности жизнью, национальных отношениях, экономической реформе.

Раздел «Повторяющиеся исследования» включает опросы мониторингового типа, проведенные ведущими социологическими агентствами. Все исследования проведены по репрезентативным российским выборкам и включают в себя повторяющийся блок вопросов, что позволяет не только проследить динамику важнейших социоэкономических показателей, но и сопоставить методики различных исследовательских институтов. Основные исследования, входящие в раздел, – это опрос экспертов и работников, «Блиц», Мониторинг ИС РАН «Зеркало мнений», еженедельный опрос «Пента», Мониторинг ИКСИ, Мониторинг социально-экономических перемен, омнибус ВЦИОМ, омнибус РОМИР, «Курьер», «Факт», «Экспресс».

В раздел «Объединенные исследования» вошли три объединенных базы данных, каждая из которых включает в себя материал, накопленный за много лет. Например, в базу «Бюджеты времени сельского населения» вошли данные за 1975–1999 гг. Это системное исследование советской и постсоветской деревни, посвященное изучение сдвигов в использовании времени при изменяющихся условиях жизни сельского населения. Исследование яв-

---

[2] Сайт Архива находится по адресу http://sophist.hse.ru/

ляется лонгитюдным относительно выборочной совокупности сельских поселений.

Компаративные исследования включают в себя «Модули ISSP» и «Модули CRRC». «Модули ISSP» – международное исследование (International Social Survey Programme), начатое в 1985 г. К настоящему моменту в нем участвует около 30 стран. Оно проводится по согласованной тематике и единой методике, что позволяет осуществлять межстрановые сравнения и отслеживать временную динамику. Россия присоединилась к странам-участницам в 1991 г. «Модули CRRC» – исследования проведенные Закавказским Ресурсным Исследовательским Центром. В настоящее время здесь находится только одно исследование – «Социально-экономическая оценка положения домохозяйств на Южном Кавказе»

Раздел «Исследования ГУ-ВШЭ» включает в себя собственные исследования, проведенные Высшей школой экономики. В частности, размещены такие проекты, как: «Исследование нерыночного сектора и структурных изменений в российской экономике», «Исследование процессов реформирования и реструктуризации промышленных предприятий», «Исследование спроса на право в области корпоративного управления».

Для удобства работы с данными, депонированными в Единый архив, разработана информационная система, позволяющая:

• найти релевантную информацию – имеется возможность поиска того или иного исследования по ключевым словам (всего ключевых слов: 1488) и по темам;

• посмотреть анкету и описание исследования;

• в режиме on-line построить линейные распределения;

• проанализировать тренды;

• получить исходные данные любого из исследований, хранящихся в Архиве.

Архив ставит перед собой несколько важных задач, которые можно определить как просветительские, формирующие стандарты исследовательского поведения и собственно исследовательскую среду.

1. Работа Единого архива направлена на то, чтобы стать центром депонирования исследований, проведенных грантополучателями различных благотворительных фондов. Депонирование данных в архив повышает ответственность исследователя, понимающего, что результаты его работы становятся публичными и контролируемыми со стороны научного сообщества, что со временем непременно скажется на уровне проводимых социологических исследований.

2. Данные Единого архива во все большей мере становятся эмпирической базой для учебных программ кафедр социальных дисциплин в университетах. И студенты, и преподаватели не только должны знать о существова-

нии архива, но и обращаться в архив при разработке учебных курсов, написании студенческих или исследовательских работ, подготовке конференций.

3. Единый архив ставит перед собой задачу стать частью европейской и международной сети архивов, быть совместимым с ними на всех уровнях — и концептуальном, и технологическом, — с тем, чтобы, с одной стороны, коллекции наших данных стали доступнее зарубежным исследователям, с другой — чтобы российские исследователи получили простой и надежный доступ к данным зарубежных коллег.

Включение данной темы в наш курс лекций не случайно и связанно, прежде всего с тем, что Единый архив хранит данные в формате SPSS и передает их для исследовательской и преподавательской работы на безвозмездной основе. Между тем, наличие качественных баз данных является обязательным условием для полноценного изучения пакета SPSS. Как показывает опыт авторов, количественные исследования, проведенные студентами самостоятельно, в большинстве случаев не соответствуют необходимым требованиям, которые позволяют качественно изучить многие статистические функции и методы анализа данных. Это является следствием ряда причин объективного характера, среди которых:

- Отсутствие ресурсов для качественного проведения полевого этапа исследования. Как следствие – тематика, объект и предмет, цели и задачи студенческих исследовательских проектов тесно связанны с доступностью респондентов. Выборка в таких случаях в основном целевая и опрашиваются такие группы, как, например: студенты, школьники, пенсионеры и т.п. Анкеты же чаще всего рассчитаны на самозаполнение, что существенно сказывается на их качестве, соответственно и на качестве исследования в целом.
- Отсутствие необходимого опыта выбора методологии и разработки методики эмпирического количественного исследования, а также отсутствие навыков обработки данных в SPSS зачастую делают результаты исследования «неуклюжими», с высокой долей привнесенного субъективизма.
- Ограниченность во времени, необходимость совмещения проведения исследования с изучением других дисциплин учебного плана и т.п.

Кроме того, существует и субъективные причины, затрудняющие получение качественного эмпирического материала, такие как: простая человеческая лень, свойственная многим студентам, а также недостаточное внимание к студенческому исследованию со стороны преподавателя – руководителя научного проекта.

Сказанное нисколько не уменьшает роли самостоятельного проведения студентами социологических исследований, но подчеркивает необходимость привлечения дополнительных методов, которые могли бы компенсировать качественные издержки, имеющие место в образовательном процессе.

Следует сказать, что рассмотренный выше архив является единым, но не единственным. Свой архив имеется и у института социологии РАН[3], который содержит результаты более чем 700 социологических исследований, проведенных и Институтом социологии РАН и другими социологическими центрами страны. Уникальность этого архива в том, что в нем хранятся данные исследований за большой период времени – примерно за 40 лет. Первое исследование датировано 1966 годом. В банке данных собраны и систематизированы материалы всесоюзных, республиканских, всероссийских и местных исследований.

Фонды архива предоставляются для использования в академических целях на безвозмездной основе. Чтобы получить данные, достаточно заполнить специальную форму-запрос и указать, для какой научно-исследовательской работы они необходимы. Форму необходимо выслать на электронный адрес. Данные также предоставляются в формате SPSS.

Приведем высказывания некоторых пользователей Единого архива, позволяющие показать его значимость.

Единый архив социологических данных открывает совершенно новые горизонты перед отечественными учеными – социологами, политологами, экономистами, социальными психологами и историками, так как дает возможность по-иному взглянуть и исследовать многие современные социальные проблемы: социальное неравенство, бедность, миграцию, доступность образования, безработицу и т. д. Это своеобразный «золотой» неиссякаемый источник, откуда можно черпать совершенно уникальную информацию. Значение Единого архива социологических данных для дальнейшего развития российской науки трудно преувеличить.

С уважением, аспирант
ИС РАН Елена Тарасенко

Создание Единого архива социологических данных – на мой взгляд, одно из самых замечательных и обнадеживающих событий в нашей науке. Когда в 70-х годах я познакомилась с работой Кельнского архива, то думала, что вряд ли доживу до создания чего-либо похожего в России. На переломе 80-х и 90-х годов аналогичный архив стал создаваться в Венгрии. Но у нас ничего подобного не было еще много лет. И вот свершилось: Единый архив социологических данных создан, его сотрудники успешно осваивают новую деятельность, ведут большую методологическую работу. Появились первые общедоступные массивы социологических данных, с которыми могут работать ученые, студенты, аспиранты. Конечно, создание полноценного архива потребует решения многих методологических, организационных,

---

[3] Архив можно найти по адресу: http://www.isras.ru/Databank.html.

финансовых и даже этических проблем. Но первые шаги в этом направлении свидетельствуют о том, что российская социология живет и успешно развивается.

Академик Т. И. Заславская

В качестве администратора данных по общественным наукам в Оксфордском университете я помогаю аспирантам и преподавателям находить и получать доступ к базам данных, необходимым для их исследовательской работы. Наличие национальных архивов данных намного облегчает эту задачу: сотрудничать с ними гораздо проще, чем заказывать информацию непосредственно у правительства, исследователей или коммерческих агентств. Недавно я получила данные для наших исследователей из Единого архива социологических данных. Я считаю работу сотрудников этого архива чрезвычайно быстрой, дружелюбной и эффективной. Данные, которые отражают изменения в российском обществе за последние годы, необычайно интересны, и мы очень рады, что они стали так легко доступны благодаря существованию и работе Единого архива.

Джейн Робертс, Оксфорд

Отдельно стоит отметить возможности Единого архива для осуществления вторичного анализа данных социологических исследований, которому уделяется слишком мало внимания в процессе обучения социологов. В настоящее время, в связи с бурным развитием информационных и коммуникативных технологий, для использования этого метода практически нет никаких препятствий.

В среде вузовских социологов можно встретить мнение, что вторичный анализ является «второсортным», по сравнению с первичным сбором данных, поскольку анализируются «старые данные», которые обладают низкой ценностью в силу отсутствия актуальности и научной новизны. Однако, по мнению авторов, такой подход является ошибочным и на самом деле вторичный анализ обладает огромным потенциалом и латентной новизной, обусловленной поверхностностью первичного анализа большинством исследователей. Более того, в настоящее время наблюдается нехватка исследований, основанных на вторичном анализе. Данная проблема уже неоднократно рассматривалась отечественными исследователями, в частности, А. В. Стрельникова указывает на то, что слабое использование вторичных данных порождает целый ряд проблем и препятствий в развитии социологической науки и научного сообщества. Среди них – неразработанность многих перспективных методов сбора и анализа социологической информации, снижение уровня культуры использования математико-статистических методов в

социологии, разобщенность исследователей и утрата преемственности знания, а также ряд других проблем.

В настоящее время вторичный анализ из довольно частного, локального и специфического метода становится все более активно используемым, широко применяемым средством получения социологического знания. Его нарастающая популярность вызвана, в первую очередь, объективными причинами. Действительно, дорогостоящая, трудоемкая, рутинная, зачастую занимающая наибольшее время и требующая наибольших организационных усилий часть эмпирического социологического исследования — это работа с респондентами, сбор первичной социологической информации или первичных социологических данных. Как правило, информативность данных, собранных специально для обоснования исходных исследовательских гипотез, проверкой, подтверждением или опровержением их справедливости не исчерпывается, они могут быть использованы повторно. Их информационная ценность возрастает, когда они используются в иных исследовательских контекстах, в сочетании с данными других эмпирических измерений. Если совокупность такого рода исследований имеет однопорядковые характеристики, то увеличивается как размер области, на которую распространяются выводы, так и временная глубина событий, к которым выводы применимы, растет обоснованность и детализация теоретических обобщений, следующих из подобного анализа. Если же корректно проведенный синтез первичных данных позволяет сочетать качественно разнопорядковую информацию, то характер обобщений и выводов выходит за пределы просто более глубокого обоснования проблемных вопросов, исходно заданных в исследовательских гипотезах, позволяет выдвигаться на более высокий уровень ассоциаций, ставить и решать проблемы более широкого методологического плана, получать нетривиальные теоретические результаты, вплоть до самых фундаментальных уровней знания.

На практике широкое распространение персональных компьютеров, информационных сетей, развитие архивов социологической информации, подготовка информационных массивов эмпирических социологических исследований в соответствии с требованиями архивных учреждений и постоянное пополнение их фондов, наряду с внедрением знаний и навыков разнообразной и достаточно изощренной обработки данных в среде непосредственно практикующих социологов вызывает значительный рост исследований, выполненных с использованием средств и методов вторичного анализа данных.

Действительно, особенностью социологической информации является многоконтекстуальная смысловая связанность, многозначность в различных системах ассоциативных коммуникаций, интерпретационных парадигмах. Статистические способы оценки связи социальных фактов обращаются к реалиям разного уровня формализации; факторы, определяемые объективно фиксируемыми характеристиками социальных ситуаций, в свою очередь, могут оказываться в роли детерминирующих в новом круге ассоциаций при анализе объектов исследования.

Таким образом, можно утверждать, что на основе даже ограниченного объема первичных данных может быть получено бесконечное множество результатов при использовании вторичного анализа этих данных. Подобный вывод приобретает реальный смысл при уже накопленных и постоянно увеличивающихся масштабах разного рода информации. Тем более, что есть большой соблазн ограничить социологическое исследование вторичным анализом, поскольку информация, инструментарий ее обработки всегда под рукой, получение нового результата гарантировано, и, при этом, исследователь избавлен от трудной, дорогостоящей и изнурительной работы по сбору информации.

Желание исчерпать информационный потенциал первичных социологических данных естественно, и результаты приносят новое знание, но вторичный анализ полезен не только этим. При многократной обработке данных возможно сосредоточить внимание на качестве первичной информации, с точки зрения применения к ней процедур фиксации формализованных и неформализованных элементов в структуре объекта и связанной с этим концептуализации исследования, стратегии дальнейшего его развития. Вместе с тем, переключение акцентов исследовательской практики преимущественно на вторичный анализ данных содержит в себе опасность формирования шаблонного подхода. На нечто подобное в экономической науке обращал внимание В. В. Леонтьев: в американских журналах исследователи в своих публикациях предпочитали обращаться к анализу моделей, а не реальной экономической информации. Аналогия хотя и не полная, но характерная. Наряду с анализом данных модельного эксперимента, анализ данных из вторичных источников при хорошей организации архивов социологической информации доступнее, сопряжен с существенно меньшим объемом рутинной работы и с большей легкостью обеспечивает «зачетный» конечный результат.

Вторичный анализ может использоваться как самостоятельный метод или как одно из средств в комплексном анализе социологических объектов. В строгом смысле слова не следует относить ко вторичному анализу всякую

попытку использовать для получения социологического знания данные, полученные в иных, независимых исследованиях, если эти знания принадлежат к иной концептуальной среде, иному теоретическому подходу. Хотя, в известной мере, при повторном использовании первичной информации происходит сближение, или даже соприкосновение, сфер вторичного и невторичного анализа. Одна и та же социальная задача может исследоваться как в рамках вторичного анализа (при этом даже создавать некоторую традицию подхода вторичного исследования проблемы), так и вне его. Параллельно могут существовать альтернативные концепции, привлекающие материал первичного исследовательского направления, но квалификация такого исследования, как принадлежащего к сфере вторичного, будет не менее значимой, поскольку в таких исследованиях самодовлеющее значение имеет иная исследовательская концепция, иной горизонт, иной срез, иной понятийный и инструментальный аппарат с иными исследовательскими задачами. Такое исследование более правомерно можно определить как независимое, хотя понятия *вторичное исследование* и *независимое исследование* кажутся сопоставимыми на некотором более высоком уровне анализа проблемы. Информация, используемая независимым исследованием, может носить обобщенный и универсальный характер. Крайней формой такого использования информации из внешних источников является широко практикуемое социологами привлечение статистических данных. Работы, использующие данные статистических бюллетеней, сборников, справочников, нельзя считать вторичными по отношению к данным источникам – они концептуально с ними практически не связаны в силу универсального характера предоставляемых ими данных.

Прежде всего надо отметить, что вторичный анализ существует во множестве версий. Поскольку формирование таких версий ничем не ограничивается, они могут возникать по самым разным поводам, вторичный анализ осознается как доступный и одновременно респектабельный способ усилить исследовательскую позицию ученого. При таком обращении ко вторичному анализу создается масса прецедентов его применения. Фрагментарность, массовость, разнообразность, высокая частота использования позволяют квалифицировать его как разновидность исследовательского подхода, имеющую статус научного движения, поскольку он не связан ни с одной предметной областью, проблемой или программой исследования и обладает абсолютной пластичностью в отношении форм и методов исследования. Это могло бы вызвать и затруднения в его определении, поскольку как метод он не обладает набором только одному ему присущих, «эндемичных» признаков.

Популярность, массовость использования как бы «укрепляют» практику применения вторичного анализа, повышают квалификацию применяю-

щих его исследователей, так как осознается существование определенного аппарата для его эффективного, корректного, убедительного применения, формулируется нормальный корпус последовательного применения.

Поскольку широкий класс исследовательских задач может решаться исключительно средствами вторичного анализа, он окончательно оформляется как научное направление. Для его эффективного функционирования может быть создана специальная инфраструктура, и она создается в виде разветвленной сети архивов социологической информации.

Практика электронного хранения и архивирования социальной информации в западных странах последовательно культивируется с 60-х гг., и с этого времени постоянно развивается и совершенствуется. Это направление организационно оформилось вместе с созданием ***архивных учреждений***, под которыми мы понимаем высокоорганизованные и технически хорошо оснащенные информационные центры, деятельность которых полностью охватывает автоматизированный процесс сбора, накопления, хранения и обработки данных. По сути, эти архивные учреждения являются интегрированными информационными центрами, располагающими огромными возможностями для привлечения на арену социальных исследований пригодной к использованию количественной информации. Среди них лидирующие источники для вторичного анализа в академической науке – Межуниверситетский консорциум политических и социальных исследований (Interuniversity Consortium for Political and Social Research – ICPSR) и архив данных Социологического исследовательского комитета (Social Science Research Commitee – SSRC), центры Ропера и Харриса в США, архивы Штайнметца в Нидерландах, Информационный центр социально-политических данных во Франции (BDSPIC.E.R.A.T), Центральный архив социальных эмпирических исследований в Германии (ZA) и другие, которые сосредоточили у себя огромный исследовательский потенциал для проведения вторичного анализа. Более того, в 1977 г. была создана Международная федерация информационных организаций для социальных наук (IFDO), которая в 1978 г. провела свою первую международную конференцию. Страны-организаторы, представленные четырьмя североамериканскими и семью европейскими архивами данных, провозгласили своей задачей координацию усилий ученых этих стран по разработке технологий в области хранения и обработки данных и, более того, в области разработки интегрированных межгосударственных баз данных для проведения сравнительных межгосударственных исследований.

Приведем примеры наиболее известных и авторитетных архивов.

Межуниверситетский консорциум политических и социальных исследований (ICPSR). Уникальный среди архивов социологических дан-

ных, архив Межуниверситетского консорциума политических и социальных исследований, расположенный на территории Университета в Мичигане, был организован в 1962 г. Целью его создания было обеспечение значительного расширения доступа исследователей к огромному массиву данных, собранных в результате проведения эмпирических исследований начиная с 1950 г. Это самое обширное в мире хранилище и служба распространения машиночитаемых социологических данных, насчитывающих несколько десятков тысяч компьютерных файлов данных по сравнительному и историческому анализу, представляющих все социальные дисциплины по более чем 130 странам мира.

Часто используемые в США первичные исследования, такие как Американское национальное исследование выборов, Панельное исследование динамики доходов и Всеобщее социальное обследование, проводились Консорциумом, поскольку важные масштабные социальные исследования всегда инициировались государством. Изучение здоровья нации, образования, потребительского поведения и занятости – все данные доступны через Консорциум. Также здесь хранятся огромные материалы из Бюро переписей, включая десятилетние периодические переписи населения, ежегодные демографические данные, начиная с 1968 г., и исследования по специальным темам. Есть в нем и неамериканские данные, относящиеся по большей части к международным вопросам и качеству жизни. Предлагаемые массивы данных охватывают широкий диапазон дисциплин, включая политическую науку, социологию, демографию, экономику, историю, образование, геронтологию, преступное правосудие, здравоохранение, внешнюю политику, и закон. ICPSR поощряет ученых во всех областях науки вносить и использовать ресурсы данных архива.

Все массивы данных в архиве ICPSR организованы в соответствии с принятой структурой, включающей в себя восемнадцать разделов. Среди них можно назвать следующие: переписи населения, экономическое поведение, конфликты, агрессия и насилие, образование, элиты и лидерство, здравоохранение, международные системы, социальные индикаторы и некоторые другие.

Деятельность Консорциума ведется по трем направлениям.

1. Консорциум хранит и распределяет машиночитаемые социологические данные, полученные и введенные собственными исследователями, а также купленные у коммерческих или государственных организаций.

2. Эта организация обучает ученых эмпирическим исследованиям посредством проведения программы «Количественные методы в социальных исследованиях», стремясь обеспечить максимально открытый доступ к данным этого архива. ICPSR поддерживает тактику равноправного доступа

участников к данным для проведения исследований и осуществления обучающих программ.

3. ICPSR вырабатывает рекомендации и методики, которые способствуют использованию продвинутых компьютерных технологий в области архивирования данных, их хранения, накопления и обработки.

Учреждения – члены ICPSR оплачивают ежегодный членский взнос, обеспечивающий им полный объем услуг, предоставляемый ICPSR. Кроме того, каждая организация, вступающая в члены ICPSR, назначает официального представителя в общий Совет архива, который координирует доступ участника к его ресурсам и выражает его интересы на общих собраниях. В настоящее время размер членских взносов варьируется от $10,5 тыс. до $2 тыс. в зависимости от категории членства. Размеры платы для каждой членской категории основаны на размере организации и типе социологической программы, которую они выполняют. Для группы организаций доступно корпоративное членство, а некоторые некоммерческие учреждения, имеющие четко определенную образовательную миссию, по решению Общего совета могут стать бесплатными членами архива ICPSR. Центральный офис ICPSR расположен в Институте социального исследования в Университете Мичигана.

Членами ICPSR являются более чем 325 колледжей и университетов в Северной Америке, а также несколько сотен учреждений, обслуживаемых зарубежными организациями в Европе, Океании, Азии, и Латинской Америке. В число зарубежных членов архива входят такие страны, как Франция, Германия, Испания, Швеция, Швейцария, Дания, Израиль, Корея, Австралия, Южная Африка, Польша и некоторые другие. Такое широкое представительство зарубежных членов в составе архива основано, прежде всего, на безусловной целесообразности членства в ICPSR: за разумную ежегодную плату участники имеют доступ к обширному архиву данных и полному диапазону его ресурсов и услуг. Кроме этого, существуют также дополнительные преимущества членства в этом архиве, которые привлекают в него все основные научные и образовательные учреждения. К ним можно отнести следующие факторы:

• ***Удобство доступа к данным***. ICPSR обеспечивает использование эффективных инструментов поиска данных, грамотное формулирование запросов о необходимой информации, которые не только оказывают содействие в быстром доступе к данным, но и предоставляют все возможности для актуализации информации, а именно: анонсируют все новые поступления по теме.

• ***Целостность и надежность данных***. Архив своей многолетней деятельностью заслужил полное доверие участников к полноте и надежно-

сти данных, так что его участники могут осуществлять самые современные исследования и решать насущные проблемы, не опасаясь за качество информации. Все данные и документация подвергаются научной экспертизе в процессе их приобретения.

• ***Совершенствование данных***. Многие из приобретенных данных ICPSR дополняет сопровождающей документацией, которая облегчает их анализ и позволяет проводить грамотные сопоставления. Некоторые массивы интенсивно обработаны (например, сняты проблемы с несопоставимостью, заполнены пустые места и т.д.), что облегчает их дальнейшее использование в исследовательской работе.

• ***Организационная структура архива*** позволяет посредством своих официальных представителей организовать целевой поиск и приобретение данных, обеспечивая удовлетворение научных интересов всех его участников.

• ***Консультации и научная экспертиза.*** Имея более чем тридцатилетний опыт в архивировании, обработке и использовании данных, ICPSR имеет возможность оказать любую поддержку ученым в методах, стратегии и решении конкретных проблем своих исследований.

• ***Рентабельность.*** Стоимость членства в ICPSR недорога по сравнению со многими другими архивами и подписками на базы данных. Кроме того, членство приносит пользу всем ученым, работающим в данном учреждении, поскольку не только предоставляет информационную базу для исследований, но и повышает их научную квалификацию.

Под покровительством ICPSR существуют два более маленьких архива: Национальный архив компьютерных данных по старению населения (NACDA), который специализируется на статистике о здоровье, пенсии, причинах смерти, стереотипах старения, и Криминальный архив (CJAIN), поддерживаемый Бюро правоохранительной статистики.

Архив данных Социологического исследовательского комитета (SSRC). Этот архив, образованный в 1967 г., – самый большой банк социологических данных в Великобритании. Его деятельность поддерживается материально Социологическим исследовательским комитетом и Университетом в Эссексе, где архив и расположен. В банке содержатся файлы данных из академических исследований, опросов общественного мнения, государственных маркетинговых исследований, рыночных обзоров. В него включено также детализированное описание текущих британских социологических обследований. Кроме этого, в архиве имеются каталоги многих других значительных хранилищ данных (как национальных, так и международных), и одной из его функций является обеспечение доступа исследователей к их данным. Архив поддерживает отношения с другими архивами и является

членом Международного консорциума политических и социальных исследований, являясь представителем Великобритании. Он также принадлежит к Международной федерации информационных организаций (IFDO) и к Международной ассоциации «Информационные услуги и технологии в социальных науках» (IASSIST).

Процесс получения данных в архиве SSRC очень прост. Исследователь должен заполнить две формы. Одна включает в себя полные требования относительно массивов данных, их типов, размерности и т.д. Вторая – согласие пользователя с определенными условиями использования данных, выдвигаемыми архивом. Данные предоставляются на магнитных носителях в формате, определенном пользователем. Копии всех шифров и оригиналы опросных листов или анкет предоставляются заказчику до того, как оформлен заказ.

Архивом публикуется два обширных каталога, предоставляющих полнейшую информацию о массивах данных и способах их получения. Прилагаются списки публикаций, связанных с этими данными, списки собственников данных и спонсоров исследования и т.д. В каталогах файлы классифицированы соответственно по двадцати трем категориям, включая социальное благосостояние, изучение общественного мнения, услуги здравоохранения, экономическое поведения и широкомасштабные долгосрочные исследования.

Архив SSRC много усилий прилагает к поиску новых первичных файлов данных. Трехгодичный «Бюллетень Архива данных» распространяет новости об услугах архива и продвижениях в области количественного социологического анализа. В нем прилагается список новых файлов данных, пополнивших архив за последнее время. Он также анонсирует конференции, встречи и семинары по использованию и обработке данных.

Архивы Штайнметца. Институт Штайнметца, названный в честь одного из основателей социологии в Нидерландах, был основан в 1964 г. Позже Институт стал архивом, который сейчас является частью Центра информации и документации социальных наук (SWIDOC). Архив предназначен для сбора, хранения и распространения социологических данных среди вторичных аналитиков. В нем собраны более 12 000 массивов данных, охватывающих все социальных науки. Все исследования различаются по природе и по масштабу, но среди них выделяются два основных:

1) исследования национальных выборов в Нидерландах;

2) серия более чем 700 еженедельных опросов общественного мнения, проводимых Нидерландским институтом общественного мнения, и рыночных исследований (NIPO) с 1962 г.

Архив Штайнметца определенным образом хранит данные, собранные из нескольких исследовательских институтов и государственных источников. Некоторые файлы данных покупаются Архивом на основе информации, получаемой из библиотеки SWIDOC и научной периодики. Исследования, сохраняемые в архиве, должны иметь основную сопутствующую документацию, включая систему кодирования, копию опросного листа, две копии о цели исследования, описание формата данных. Могут быть отмечены конкретные ограничения использования данных.

Архив систематически классифицирует каждое исследование в своем «Каталоге» и «Путеводителе» соответственно международной схеме, которая требует обозначения идентификационного номера исследования, заголовка, даты начала, ключевых слов, числа переменных и т.д. Каталожные индексы позволяют исследователю расположить массивы данных по ключевым словам, названию, дате, руководителям проектов.

Архив предлагает пользователю множество услуг, некоторые бесплатно. Получить данные из архива довольно просто. Достаточно заполнить форму, напечатанную на обороте каталога. Данные обычно предлагаются в машиночитаемом виде, чаще всего в файлах SPSS.

Широко распространена практика чтения работниками этого архива обучающих лекций, проведения семинаров по новейшей обработке данных, по созданию специфических файлов данных из собранных материалов, а также по вторичному анализу.

Архив Штайнметца является членом Международной федерации информационных организаций (IFDO), Международной ассоциации социологических информационных услуг и технологий (IASSIST), а также членом ECPR и ICPSR.

Основные архивы опросов общественного мнения: Центр по изучению общественного мнения Ропера и Центр данных Льюиса Харриса. Данные опросов общественного мнения являются богатым источником для вторичного анализа. Основные центры по изучению общественного мнения в США — центр Ропера и Центр Харриса.

Основанный в 1946 г., **центр Ропера** по изучению общественного мнения стал самым большим по теме архивом в мире. Центр не является в прямом смысле архивом данных опроса, поскольку некоторые академические исследования (например, Всеобщее социологическое исследование) проводились с участием сотрудников этого центра. Начиная с 1977 г. центр Ропера обосновался в Университете Коннектикута, который управляет Центром наряду с Иельским Университетом и колледжем Уильямса. Центр имеет более чем 10 000 файлов данных и каждый год добавляет по 500 новых файлов. Обследования проводились начиная с 1930 г. вплоть

до настоящего времени и охватывают Соединенные Штаты и еще более 70 государств. В основном массивы данных Центра относятся к социальным показателям, социальным и политическим предпочтениям, личностным оценкам.

Многие организации, специализирующиеся на обследованиях (например, Американский институт общественного мнения, Национальный центр по исследованию общественного мнения и т. п.) регулярно пополняют массивы данных центра Ропера. Результаты многих известных специфически социологических исследований хранятся в базе данных этого Центра. Среди них можно назвать исследование Самуэля Стаффера «Обследование американских солдат Второй мировой войны», проект Алекса Инкеля «Вхождение в современность», посвященный изучению социальных и культурных аспектов процесса развития, обследование мнений американских женщин, серия исследований «Состояние нации» и различные массивы данных, описывающих американский электорат с различных точек зрения.

Для получения данных из Центра необходимо стать его членом. Центр предлагает три категории членства: для колледжей, для университетов и для внешних организаций. За небольшие по размерам ежегодные взносы эти организации получают доступ к определенным массивам данных (есть и некоторые ограничения) и к соответствующим услугам. Дополнительные данные и услуги доступны за дополнительную плату, причем она существенно ниже, чем для организаций, не имеющих членства в Центре.

Центр Ропера выпускает несколько периодических публикаций, например, «Обзор данных для прогнозирования» с описанием обследований, регулярно проводящихся с 1930 г., бюллетени новых данных и еще несколько очень ценных изданий.

Центр Льюиса Харриса, основанный в 1965 г., сейчас является частью Библиотеки данных по социальным наукам при Университете Северной Каролины. «Льюис Харрис и компаньоны» — частная международная фирма, которая с 1956 г. проводила обследования, в том числе непрерывное исследование «ABC News/Harris». Фирма «Харрис и К» и Университет в Северной Каролине создали огромный архив данных, куда вошли результаты исследований фирмы, проводящиеся начиная с 60-х годов. В свою очередь, Университет взял на себя ответственность за организацию хранения и распространения данных.

Центр Харриса содержит сотни файлов данных, затрагивающих такие интересные проблемы, как подвижки в статусе женщин, налогообложение, выборы, контроль за вооружением, покупательские предпочтения. Причем

выборки проводились как из специфических популяций (например, врачей, путешественников через Атлантику и государственных лидеров), так и из общих популяций (население в целом, электорат).

Вышеприведенный анализ некоторых архивов данных свидетельствует о том, что ключевая проблема для вторичного анализа — выбор адекватного массива данных — в настоящее время решается без существенных проблем. Все архивы публикуют каталоги и путеводители по своим базам данных со всей необходимой сопутствующей документацией. Достаточно только четко сформулировать свои требования к информации, область своих интересов и провести дополнительные консультации с программистами выбранного архива.

Создание европейской базы данных. В заключение лекции следует остановиться на некоторых новых тенденциях, которые характерны для современного мира. Развитие национальных архивов позволило перейти к осуществлению идеи создания объединенных межгосударственных баз данных.

Многое из того, что 30 лет назад казалось неосуществимым, удалось реализовать. Созданы или создаются хранилища баз данных социальных наук в большинстве западноевропейских стран. С конца 60-х гг. регулярно проводятся международные семинары по данной тематике, в которых принимают участие исследователи со всей Европы и представители других континентов. Получили свое развитие инструментарии и методы, используемые в сравнительных исследованиях. Значительно укрепился фундамент для проведения профессиональных исследований по мере развития в Европе широкой координации в коммерческом секторе. В проводимых исследованиях широко используются телекоммуникации, компьютерные сети и новейшие компьютерные программы.

Группа европейских архивов содержит разнообразные данные: обзорную информацию, статистическую макроинформацию, совокупные базы данных, сведения на региональном уровне, а также текстовую информацию с 1944 г. Некоторые архивы содержат также исторические сведения предшествующих веков. В хранящихся материалах отражены почти все аспекты социальной жизни.

Архивы социологических данных в настоящее время имеются во многих европейских странах. На рис.1.2 приведена карта Европы с существующими в разных странах архивами данных, являющимися членами и кандидатами в члены Совета европейских архивов социологических данных. Все они имеют связи с высшими учебными заведениями, многие финансируются национальными советами по социологическим исследованиям.

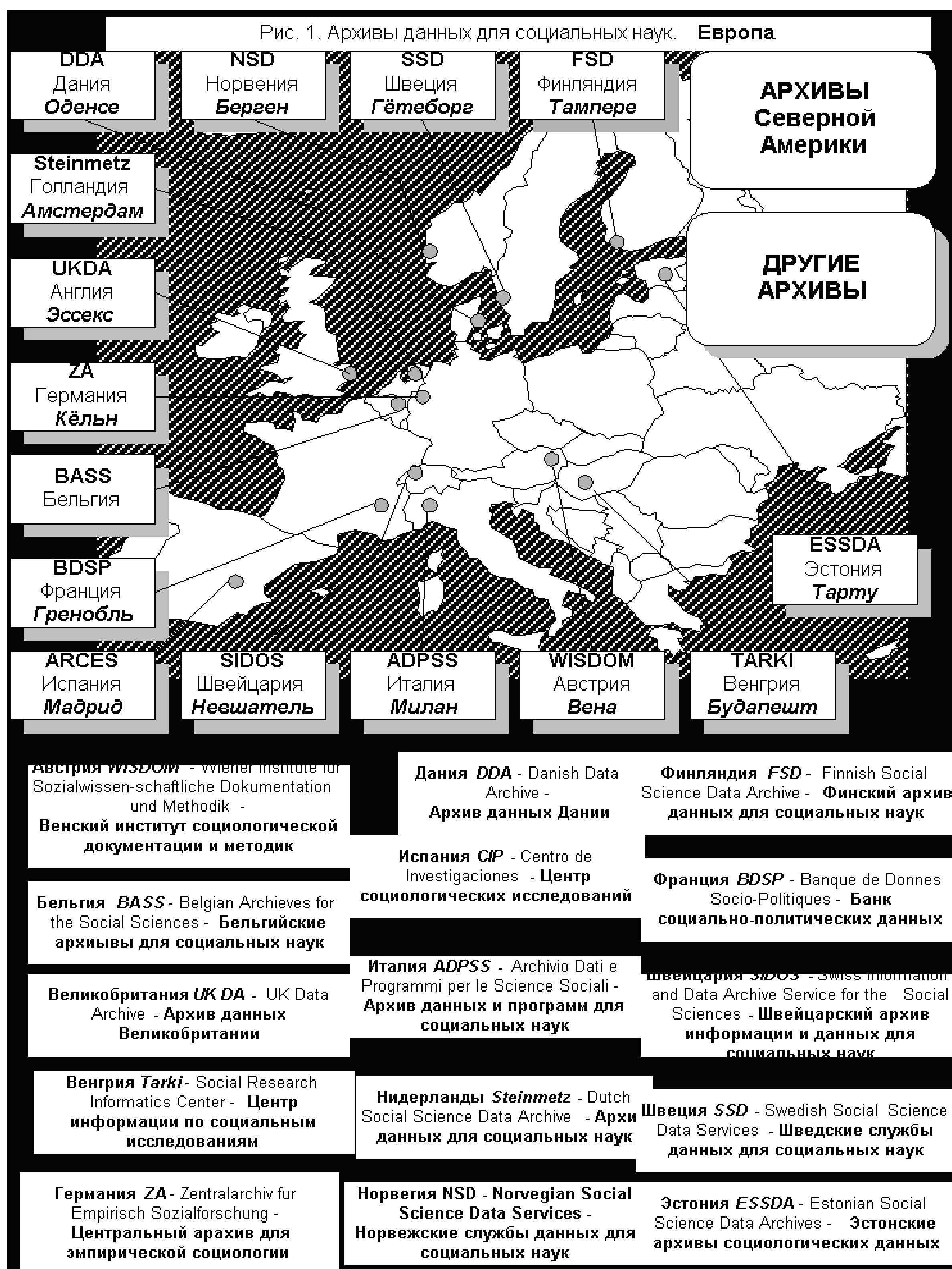


Рис. 1.2. Архивы данных для социальных наук. Европа

Вопросы и задания

1. Для чего необходимы архивы социологических исследований?
2. Зайдите на сайт Единого архива экономических и социологических данных, выберите какое-нибудь одно социологическое исследование и опишите его (кто и когда его проводил, какого характера вопросы задавались респондентам и т.д.)
3. Что вы понимаете под термином «Вторичный анализ»?
4. В каких сферах современной жизнедеятельности мы можем использовать вторичный анализ данных?

Список литературы

1. Единый архив социологических данных — новый информационный ресурс для исследовательской и преподавательской работы//Социальная реальность. — 2006. — № 9. — С. 91–95.

2. Хахулина, Л. А. Создание единого архива социологических данных: проблемы и перспективы/Л. А. Хахулина, Л. Б. Косова//SPERO — 2007. — № 6. — С. 203–208.

3. Сайт единого архива экономических и социологических данных [Электронный ресурс]. URL: http://sophist.hse.ru.

4. Сайт Института Социологии Российской академии наук [Электронный ресурс]. URL: http://www.isras.ru.

5. Стрельникова, А. В. Информационный ресурс эмпирического исследования (Проблема неполного использования): автореф. дис. … канд. социол. наук/А. В. Стрельникова. — М.: Ин-т социологии РАН, 2005. — 24 с. ргб

6. Стрельникова, А. В. Исследовательские архивы: расширение возможностей для вторичного анализа/А. В. Стрельникова//Социологические исследования. — 2005. — № 1. — С. 126–131.

7. Погорецкий, В. Г. Системные исследования. / В. Г. Погорецкий // Методологические проблемы. Ежегодник — 1998. — Часть II.

8. Давыдов, А. А. Социальная информатика. Основания. Методы. Перспективы/А. А. Давыдов, А. Б. Бритков, Т. И. Жукова. — М.: Либроком, 2010. — 216 с.

Лекция 2. Первое знакомство с SPSS 17. Подготовка матрицы и внесение данных в программу

Прежде чем переходить непосредственно к SPSS, отметим, что в настоящее время в России наиболее распространенными операционными системами являются продукты Microsoft Windows, в частности, сюда можно отнести Windows XP, Windows Vista, которые устанавливаются на большинстве компьютеров и ноутбуков. Большую популярность завоевывает и относительно недавно вышедший в свет Windows 7. В связи с этим программное обеспечение компании SPSS до недавнего времени выпускалось в основном под операционные системы корпорации Microsoft.

Не будем заострять внимания на вопросах, связанных с инсталляцией программного обеспечения SPSS, поскольку с каждым лицензионным диском идет подробная инструкция по установке программы, и, кроме того, установщик имеет удобный и доступный интерфейс, который пошагово поможет установить пакет.

С запуском программы также не должно возникнуть особых сложностей, поскольку это стандартная процедура, которую можно осуществить через меню «Пуск», выбрав пункт «Все программы» и найдя там SPSS. Кроме того, запуск можно осуществить непосредственно двойным щелчком по соответствующему ярлыку, который при инсталляции по желанию пользователя может быть размещен на рабочем столе компьютера.

При запуске программы появляется окно редактора данных, которое в левом нижнем углу имеет две вкладки: «Данные» и «Переменные» (рис. 2.1). Во вкладке «Переменные» кодируется анкета, по которой было проведено социологическое исследование, а вкладка «Данные» нужна для переноса в нее ответов респондентов со всего собранного массива анкет. Переход между вкладками осуществляется простым щелчком мыши по нужной вкладке.

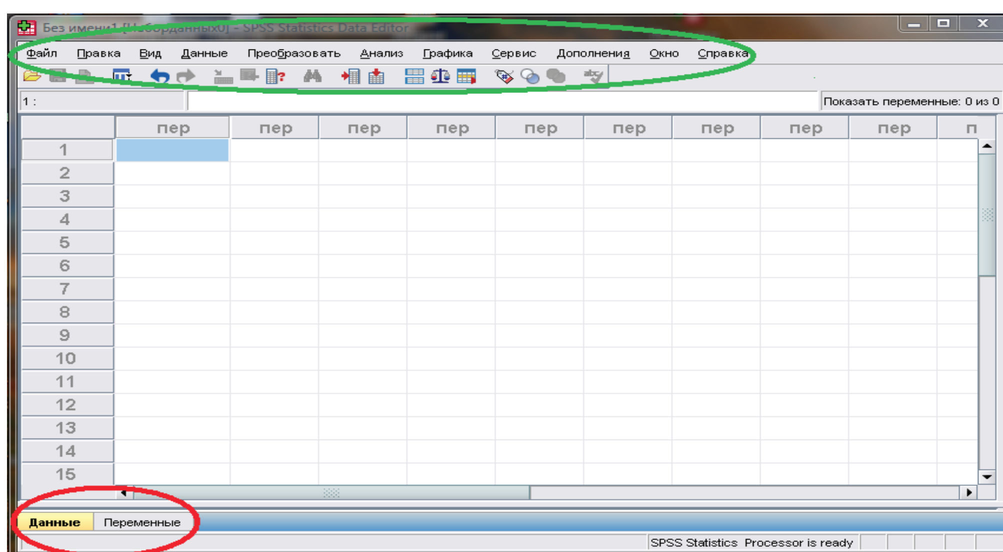


Рис. 2.1. Окно редактора данных SPSS 17

В верхней части окна находится меню, с помощью которого осуществляется весь процесс работы с данными, однако сейчас мы не будем рассматривать подробно все его пункты, чтобы не перегружать читателя лишней информацией. Вернемся к ним в процессе работы в SPSS позже, по мере возникновения в этом необходимости.

Для логического завершения первого знакомства с интерфейсом программы отметим, что при осуществлении какого-либо действия в меню программы появляется второе окно – окно вывода, в котором фиксируется вся история операций, производимых пользователем в процессе работы. Выглядит окно вывода следующим образом (рис 2.2):

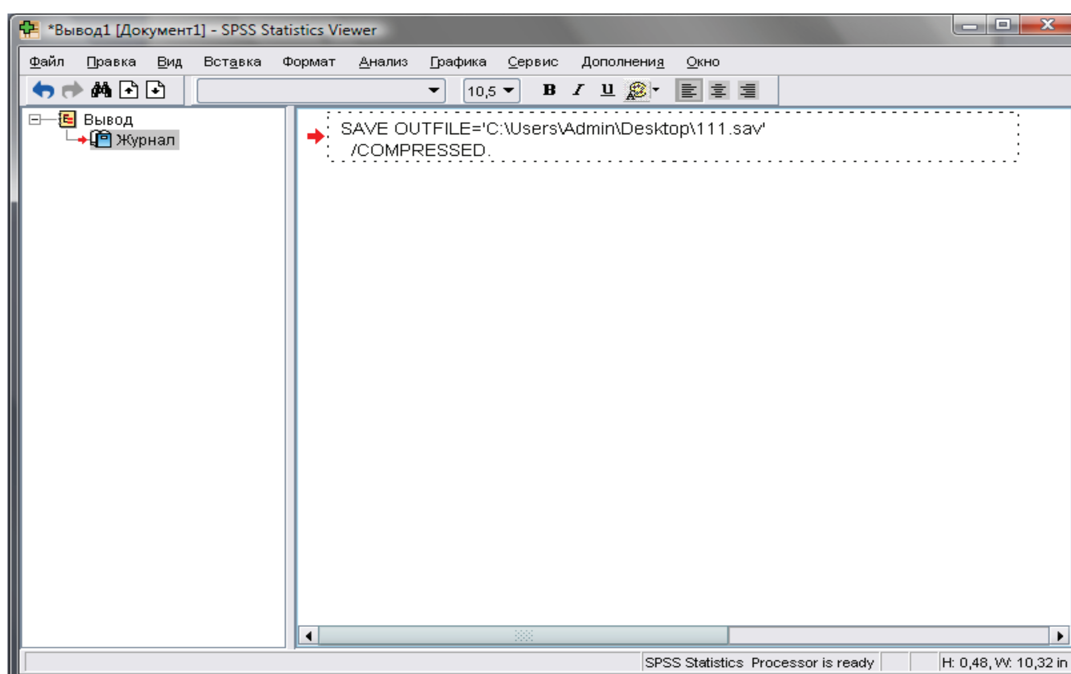


Рис. 2.2. Окно вывода

Итак, приступим непосредственно к работе в SPSS. Допустим, имеется массив анкет, полученных в результате социологического исследования, которые необходимо обработать в программе. С чего начать? Начинать нужно с составления матрицы анкеты в программе. Это своего рода «скелет», состоящий из вопросов анкеты, на который впоследствии будет наращиваться «тело» в виде многочисленных ответов респондентов.

Матрица составляется во вкладке «Переменные» окна редактора данных, которая выглядит следующим образом (рис.2.3). Каждая строка пронумерована и предназначена для создания только одной переменной. На рисунке видно, что в SPSS переменные могут иметь девять основных параметров, представленных соответственно девятью столбцами. Рассмотрим каждый из них подробнее.

1. Имена переменных нужны для того, чтобы программа могла различать их. Именно поэтому имя каждой переменной должно быть уникаль-

ным, дублирующиеся имена не допускаются. Существуют и другие требования, в частности, имена переменных могут иметь длину до 64 символов, первый из которых должен быть буквой либо одним из символов – @, #, или $. Последующие символы могут быть любой комбинацией букв, чисел, точек и др. Кроме того, имена переменных не должны содержать пробелов.

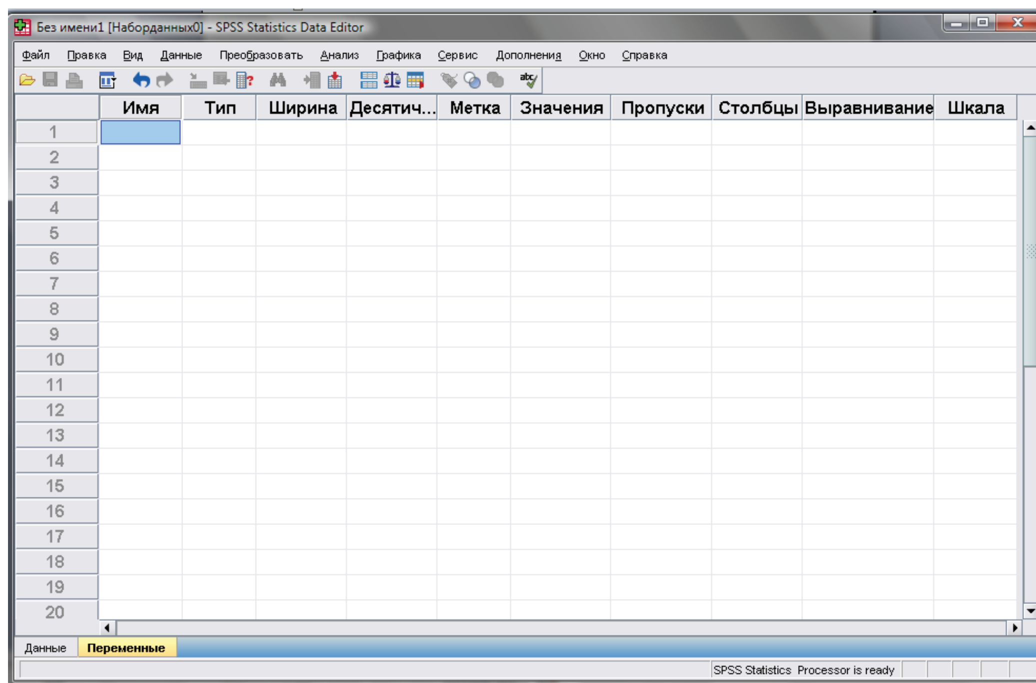


Рис. 2.3. Вкладка «Переменные» Окна редактора данных

2. Тип переменной позволяет указать тип данных для каждой переменной. Если нажать левой кнопкой мыши в ячейке, которая находится в столбце «Тип», можно увидеть, что в этой ячейке появляется небольшая кнопка (рис. 2.4), нажав на которую активизируется диалоговое окно[4] (рис. 2.5).

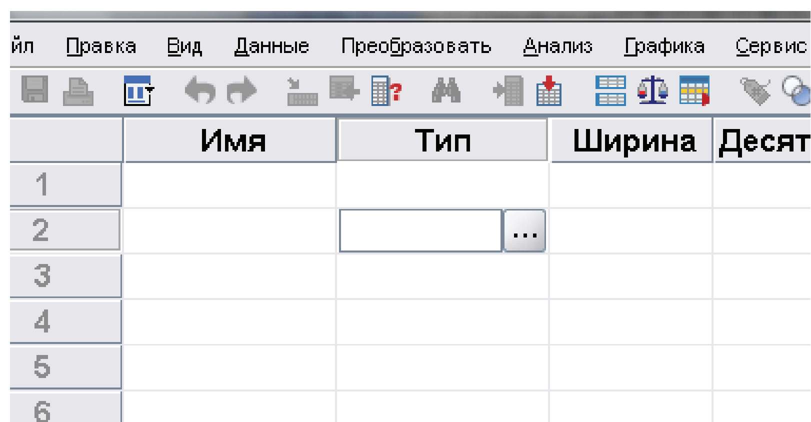


Рис. 2.4. Ячейка «Тип» после нажатия на ней левой клавишей мыши

[4] Отметим, что подобным образом активируются все диалоговые окна при создании матрицы.

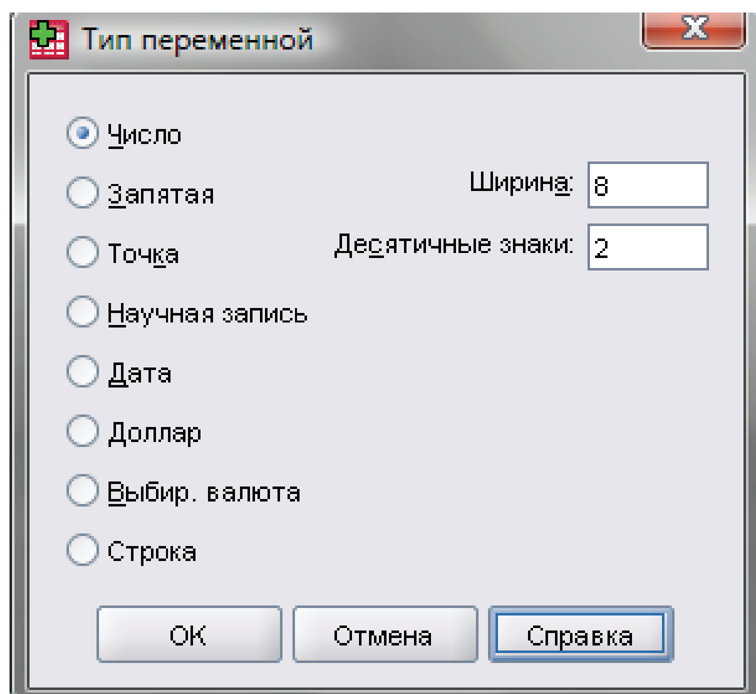


Рис. 2.5. Диалоговое окно «Тип переменной»

В появившемся диалоговом окне задается тип данных для каждой переменной. Для некоторых типов данных появляются поля для ввода ширины переменной и числа знаков после запятой, для других можно просто выбрать формат из списка с форматами.

В SPSS 17 доступны следующие типы данных (рис. 2.6):

- **Числовая.** Переменная, значения которой являются числами. Значения отображаются в стандартном числовом формате. При вводе данных Редактор данных принимает числовые значения в стандартном формате или в научной записи.
- **Запятая.** Числовая переменная, значения которой отображаются с запятыми, разделяющими каждые три разряда, а для отделения дробной части используется точка. В значениях не могут содержаться запятые справа от десятичного разделителя.
- **Точка.** Числовая переменная, значения которой отображаются с точками, разделяющими каждые три разряда, а для отделения дробной части используется запятая. В значениях не могут содержаться точки справа от десятичного разделителя.
- **Научная запись.** Позволяет задать числовую переменную, значения которой выводятся с показателем степени, представленным буквой «Е», за которой идет знак и величина степени десятки.
- **Дата.** Числовая переменная, значения которой отображаются в одном из нескольких форматов календарной даты или времени.

Формат выбирается из списка. Разделителями могут быть слэши[5], дефисы, точки, запятые или пробелы.

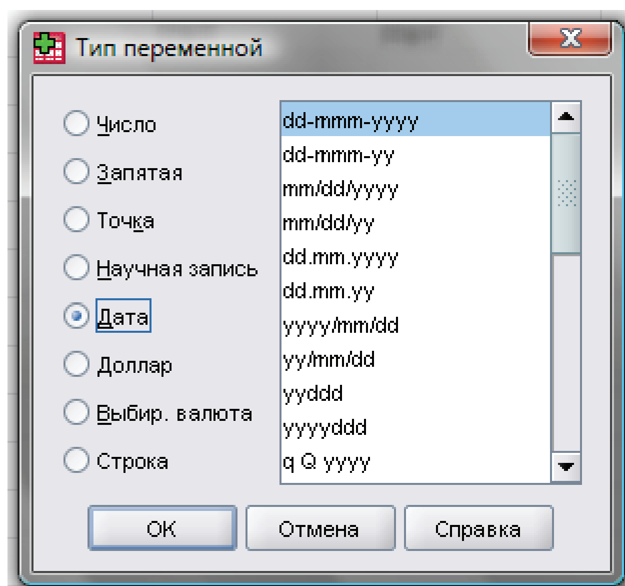


Рис. 2.6. Пункт «Дата» диалогового окна «Тип переменной»

- **Доллар.** Числовая переменная отображается со значком доллара вначале ($), точками, отделяющими группы по три разряда, и точкой в качестве десятичного разделителя. Значения данных можно вводить как со знаком доллара вначале, так и без него.
- **Выбираемая валюта.** Числовая переменная, значения которой выводятся в одном из денежных форматов, заданного пользователем на вкладке «Валюта» диалогового окна «Параметры» в разделе меню «Правка». Заданные символы валюты нельзя использовать при вводе данных, однако они выводятся в Редакторе данных.
- **Текстовая.** Переменная, значения которой не являются числовыми, не может использоваться в вычислениях. Текстовая переменная может содержать любые символы, однако их число не должно превышать величину, заданную при выборе этой переменной. Как правило, текстовая переменная необходима для кодировки открытых и полузакрытых вопросов анкеты.

Опыт авторов показывает, что в реальной практике научных социологических исследований используются в основном числовые и текстовые

[5] Косая черта (/) (в информатике – слеш, от англ. *(forward) slash*, в номерах и индексах – дробь) – типографский знак в виде тонкой прямой линии, наклонной вправо. Обычно изображается несколько выступающим вверх и вниз за линию прописных букв и цифр шрифта. Источник: Википедия – свободная интернет-энциклопедия (http://ru.wikipedia.org/).

переменные, несколько реже «Дата»[6]. Можно предположить, что в маркетинговых исследованиях этот набор шире, однако в рамках данной работы особое внимание будем уделять только указанным типам переменных.

3. Ширина – параметр, позволяющий ограничить количество символов, вводимых в ячейку. Это необходимо для удобства работы с данными. Логично, что текстовые переменные будут иметь гораздо большую ширину, чем числовые.

4. Десятичные. Этот параметр ограничивает количество символов после запятой, если вводятся десятичные числа.

5. Метка переменной. По своей сути, это вопрос анкеты. Если вопрос в анкете сформулирован длинно и сложно, то целиком внести его в программу не удастся, поскольку SPSS позволяет создать метку переменной только длиной до 256 символов. Метки переменных могут содержать пробелы и любые другие символы, которые не допускается применять в именах переменных.

6. Значения по своей сути представляют собой подсказки к вопросу, которые может выбрать респондент, если вопрос не носит открытый характер. Каждому значению переменной можно присвоить содержательную метку, например, коды 1 и 2 для обозначения пола мужской и женский соответственно (рис. 2.7). Метки значений могут быть длиной до 120 символов.

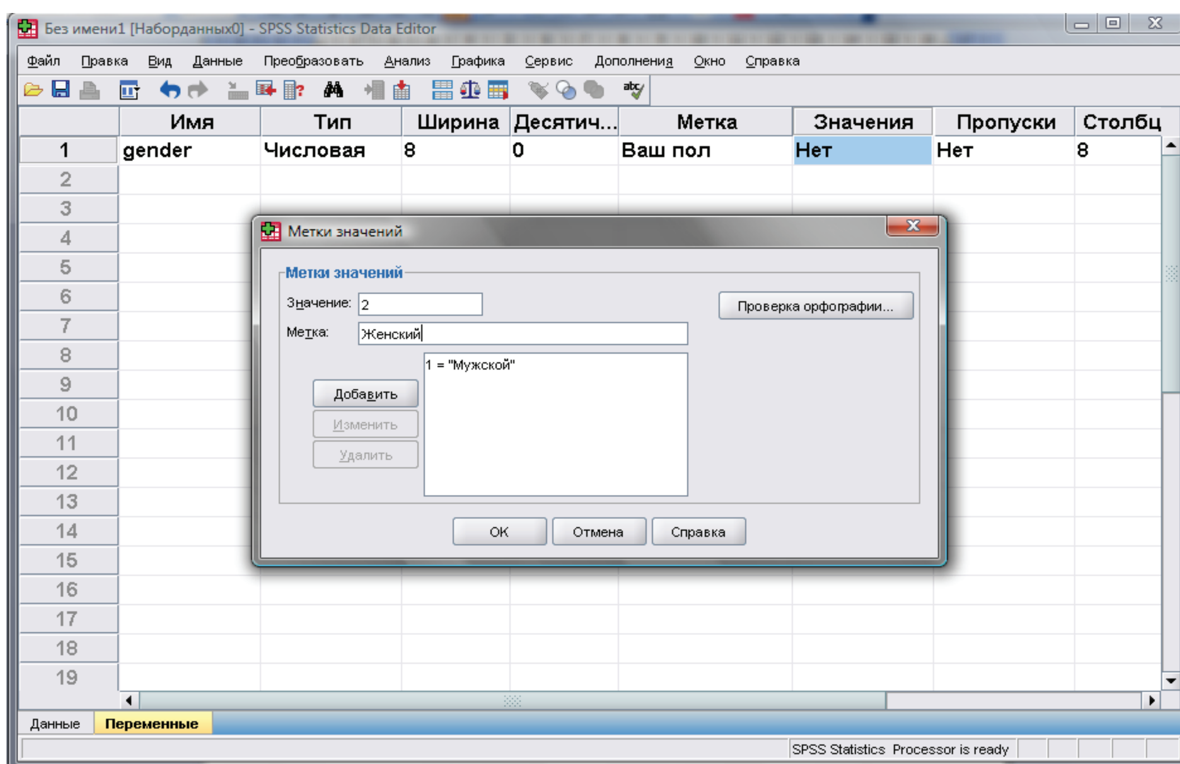


Рис. 2.7. Метки значений

[6] В том случае, если по замыслу разработчиков анкеты респонденты указывают, например, свой возраст не в виде количества полных лет, а в виде даты своего рождения.

7. Пропуски. В диалоговом окне «Пропущенные значения» определенные значения задаются как пользовательские пропущенные. Например, необходимо отделить данные, пропущенные из-за отказа респондента отвечать, от данных, пропущенных из-за того, что вопрос не относится к респонденту. Значения данных, обозначенные как пользовательские пропущенные, помечаются для специальной обработки и исключаются из большинства вычислений.

В программе имеется возможность введения до трех отдельных пропущенных значений, диапазон пропущенных значений или диапазон плюс одно отдельное значение (рис. 2.8). Диапазоны пропущенных значений могут быть заданы только для числовых переменных, а все текстовые значения, включая пробелы и пропуски, считаются валидными.

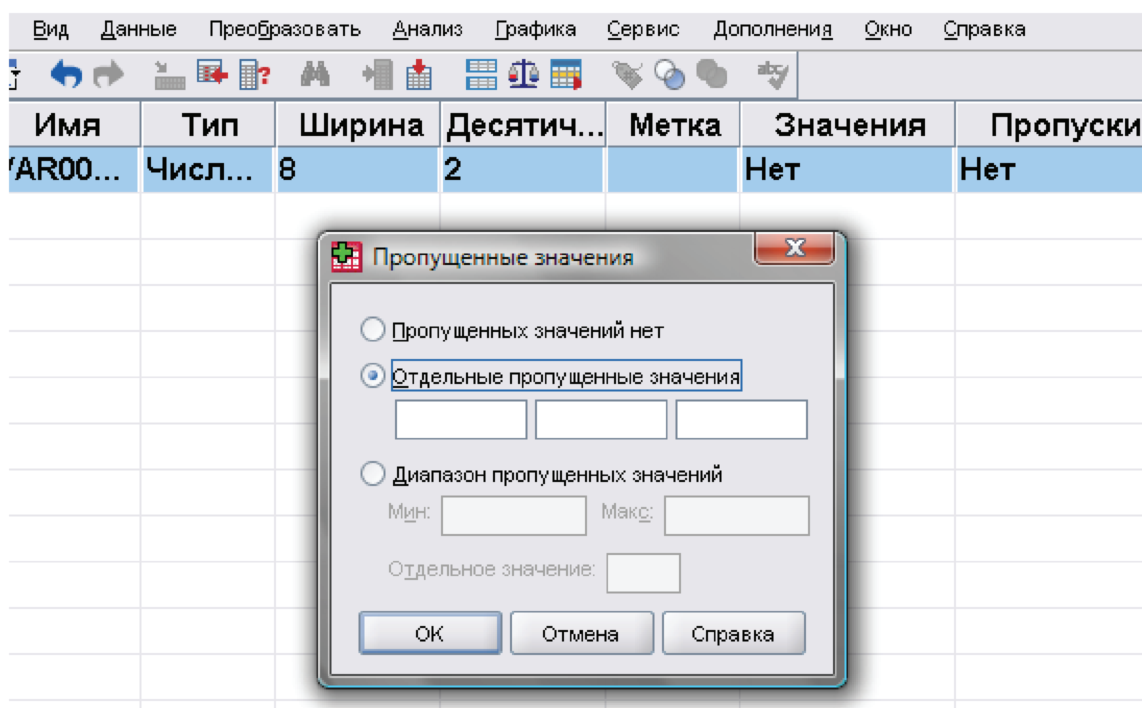


Рис. 2.8. Пропущенные значения

8. Столбцы. При помощи этого параметра регулируется количество символов, задающих ширину столбца в Редакторе данных. Ширину столбца можно также изменить в Редакторе данных на вкладке «Данные», перетащив мышью правую границу столбца. Ширина столбца влияет лишь на представление значений в Редакторе данных. Изменение ширины столбца не изменяет заданной ширины переменной.

9. Выравнивание переменной позволяет изменить местоположение данных в окне Редактора данных. По умолчанию числовые переменные выровнены по правому краю, а текстовые переменные – по левому.

Выравнивание влияет только на представление (внешний вид) данных в Редакторе данных.

10. Шкала измерения переменной позволяет задать шкалу измерения переменной: количественную, порядковую, или номинальную (рис. 2.9).

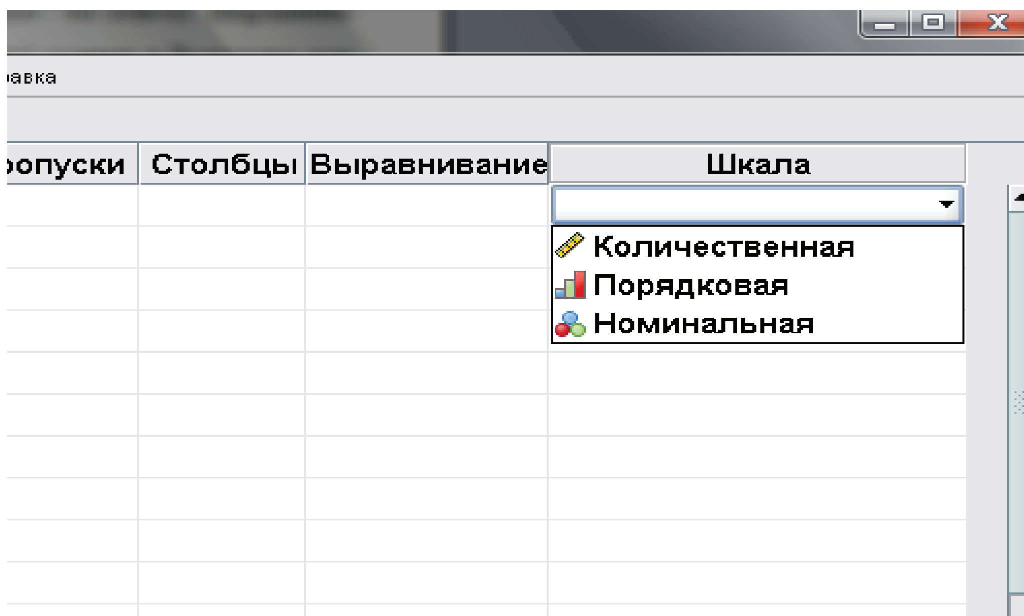


Рис. 2.9. Шкалы измерения переменной

- Количественная шкала представляет собой числовые данные с интервальным (возраст респондентов в категориях: 18–25 лет, 26–30 лет, 31–35 лет и т. д.) или абсолютным уровнем измерения (возраст респондентов, измеряемый в количестве полных лет). Переменную можно рассматривать как количественную, когда ее значения представляют упорядоченные категории с осмысленной метрикой, так что уместно сравнивать расстояния между значениями.

Номинальные и порядковые данные могут быть текстовыми (алфавитно-цифровыми) или числовыми.

- Номинальные. Переменную можно рассматривать как номинальную, когда ее значения представляют категории без естественного упорядочения, например, подразделение компании, где работает наемный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религиозную конфессию.
- Порядковые. Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности обслуживанием – от крайней неудовлетворенности до крайней удовлетворенности. Примеры порядковых переменных включают

баллы, представляющие степень удовлетворенности или уверенности, или баллы, оценивающие предпочтение.

У читателя резонно возникает вопрос: для чего нужно деление шкал на несколько видов? В социологическом исследовании эта тема является одной из основных, поэтому позволим себе сделать небольшой экскурс в проблему.

Дело в том, что ряд социальных свойств человека, такие, например, как возраст, уровень заработной платы, количественно определены, однако большинство социальных явлений и процессов такой количественной определенности не имеют. К ним относятся эмоциональные и поведенческие акты, а также суждения и мнения людей по различным вопросам. Для социолога важно определить не только их наличие или отсутствие, но также интенсивность их проявления. Чтобы решить эту задачу, при проведении эмпирического исследования социолог вынужден создавать специальную процедуру приписывания количественной определенности изучаемым качественным признакам. Такая процедура называется измерением.

Инструментом измерения выступает шкала. С помощью шкал могут быть изменены почти все, даже самые сложные, социальные явления. Шкала представляет собой систему характеристик изучаемого свойства, выполняющую роль эталона. С целью разработки шкалы определяют крайние состояния изучаемого процесса или явления – начало и конец, максимум и минимум. При нахождении крайних точек устанавливается дробность шкалы с помощью делений.

Существование различных видов шкал обуславливается необходимостью приведения качественно разнородных данных к сопоставимым количественным показателям. В программе SPSS можно задать следующие шкалы:

- **Номинальная.** Переменную можно рассматривать как номинальную, когда ее значения представляют собой категории без естественного упорядочения, например, регион проживания респондента, почтовый индекс или религиозная конфессия.
- **Порядковая.** Переменную можно рассматривать как порядковую, когда ее значения представляют категории с некоторым естественным для них упорядочением, например, уровни удовлетворенности деятельностью органов власти: от крайней неудовлетворенности до крайней удовлетворенности.
- **Количественная.** Переменную можно рассматривать как количественную, когда ее значения представляют упорядоченные категории с осмысленной метрикой, так что уместно сравнивать расстояния между значениями. Примеры количественной переменной включают возраст в годах и доход в рублях.

Отметим, что номинальные и порядковые данные могут быть текстовыми (алфавитно-цифровыми) или числовыми.

Рассмотрим на конкретном примере, как кодируются различные вопросы анкеты в SPSS. С этой целью была создана небольшая «сборная» анкета, включающая в себя различные по характеру вопросы из различных исследований, в которых принимал участие автор.

Естественно, что приведенная анкета не может включить в себя все разнообразие типов вопросов, которые имеют место вследствие опыта различных исследователей и задачами, которые он преследует при составлении социологического инструментария. Однако, на наш взгляд, рассмотрение приведенных примеров позволит уловить общие принципы и алгоритм кодирования вопросов в SPSS, что может существенно облегчить задачу и проявить творчество при кодировании особо «необычных» вопросов.

АНКЕТА

1. Ваш пол

1. Мужской
2. Женский

2. Ваш возраст (укажите количество полных лет)_______

3. На кого, по-вашему, ориентируются современные студенты в выборе жизненных стратегий, стиля жизни, в повседневности? (отметьте три наиболее важных позиции):

1. Литературные герои
2. Политики, государственные деятели
3. Бизнесмены, успешные и богатые современники
4. Преподаватели вузов
5. Деятели культуры, искусства
6. Звезды шоу-бизнеса
7. Известные ученые
8. Светская «тусовка»
9. Звезды телевидения
10. Родители, родственники
11. Друзья, знакомые из молодежной среды
12. Другое (укажите свой вариант)___________________________

4. Оцените, как часто на лекционных и семинарских занятиях в вашем вузе поднимаются следующие темы.

Темы	Часто	Иногда	Поднимали 1-2 раза	Практичес-ки не поднимаются	Не задумывался над этим
1. Профессиональная этика	1	2	3	4	5
2. Терпимость к людям другой веры, национальности и т.п.	1	2	3	4	5
3. Деятельность институтов гражданского общества (общественных организаций, партий и т.д.)	1	2	3	4	5
4. Способы и возможности самоорганизации и самоуправления молодежи	1	2	3	4	5
5. Личная гражданская ответственность	1	2	3	4	5
6. Духовные, нравственные основы жизни человека	1	2	3	4	5
7. Проблемы молодежной девиации (наркомания, алкоголизм, преступность)	1	2	3	4	5

5. На какой ступеньке «сидят» сегодня большинство студентов и преподавателей в шкале «вечных» ценностей (обозначьте: С – студенты, П – преподаватели)?

1. Лестница честности	2. Лестница справедливости	3. Лестница патриотизма	4. Лестница толерантности

Начнем с первого вопроса – о поле респондента. Он носит закрытый характер и предполагает всего два варианта ответа «1. Мужской» и «2. Женский». Кодировка этого вопроса в SPSS осуществляется в окне Редактора переменных следующим образом:

1. Сначала присваиваем имя переменной, допустим, «gender»;
2. Выставляем тип переменной – «числовой»;
3. Ограничиваем количество символов после запятой в столбце «десятичные» до «0», т.к. при внесении данных из анкет в программу у нас будут использоваться только числа «1» и «2», соответствующие двум вариантам ответа – «мужской» и «женский», при этом никаких дробных значений использоваться не будет;
4. В столбце «ширина» указываем значение «1», т.к. числа 1 и 2 являются одноразрядными;

5. В качестве «метки» вносим суть задаваемого вопроса, то есть «Пол респондента»;
6. В «значениях» вносим варианты ответов, предложенные респондентам, то есть «мужской» и «женский», а также дополнительный вариант 99 – «Нет ответа»[7] (рис. 2.10);

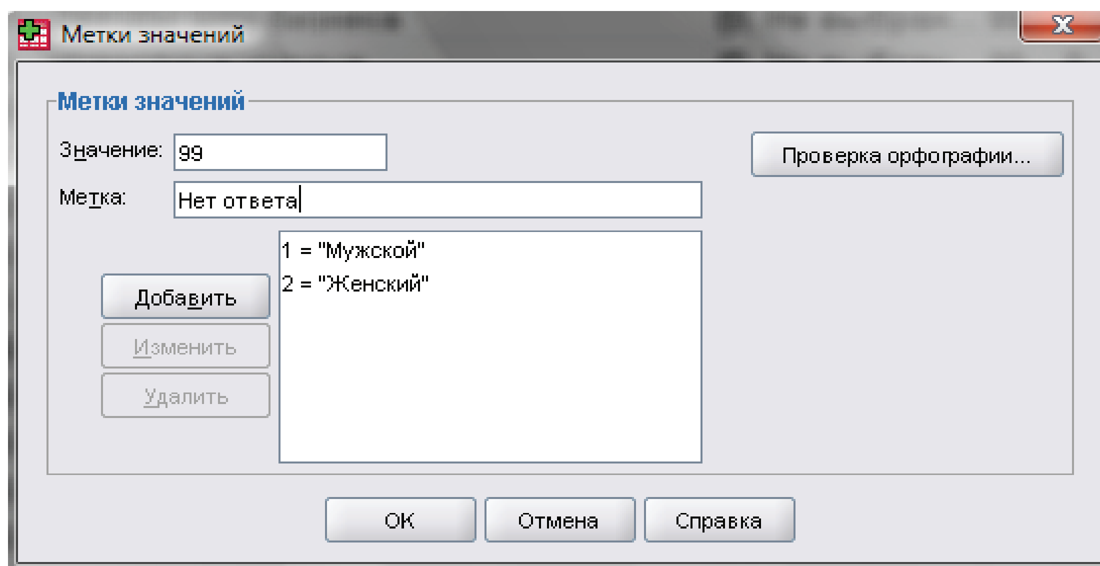


Рис. 2.10. Кодирование вариантов ответов на вопрос о поле респондента

7. В «пропусках» вызываем диалоговое окно «Пропущенные значения», выбираем пункт «Отдельные пропущенные значения» и вносим в первую ячейку число «99» (рис. 2.11). В дальнейшем при «забивке» анкет, если респондент не указал свой пол (в случае, если анкета рассчитана на самозаполнение), будем вносить в ячейку именно число 99. Отметка пропущенных значений играет существенную роль при кодировании переменных. Она позволяет отделить данные, пропущенные из-за отказа респондента отвечать, от данных, пропущенных из-за того, что вопрос не относится к респонденту. Значения данных, обозначенные как пропущенные, помечаются программой для специальной обработки и исключаются из большинства вычислений. SPSS позволяет ввести до трех отдельных пропущенных значений, диапазон пропущенных значений или диапазон плюс одно отдельное значение. Диапазоны пропущенных значений могут быть заданы только для числовых переменных. Чтобы для текстовой переменной задать пустые значения или пробелы как пропущенные, необходимо ввести одиночный пробел в одно из полей для отдельных пропущенных значений;

[7] Число взято условно, может использоваться любое другое, по желанию исследователя

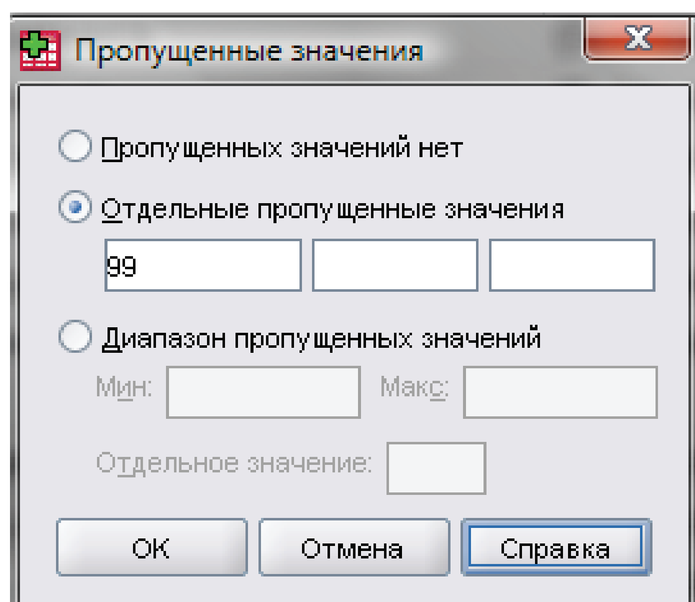


Рис. 2.11. Диалоговое окно «Пропущенные значения»

8. В разделах «Столбцы» и «Выравнивание» выставляем удобные для нас параметры. В нашем примере это «8» (ширина столбца в окне Редактора данных, измеряемая в количестве символов) и «По правому краю» (место расположения данных в ячейке) соответственно;
9. В конце выставляем «Шкалу», которая в нашем случае является номинальной.

В итоге у нас получилась переменная, позволяющая занести в программу пол респондентов со всего массива анкет (рис. 2.12).

Файл Правка Вид Данные Преобразовать Анализ Графика Сервис Дополнения Окно Справка

	Имя	Тип	Ширина	Десятичные	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	gender	Числовая	1	0	Пол респондента	{1, Мужской}...	99	8	По право...	Номинальн...
2										
3										
4										
5										

Рис. 2.12. Кодирование вопроса о поле респондента

Одновременно с созданием очередной переменной в окне редактора данных появляется новая колонка, в которую с каждой анкеты вносятся свои данные (рис. 2.13).

При забивке в программу данных анкет о поле респондентов в соответствующую ячейку будут вноситься только численные значения – «1» или «2». Второй вопрос анкеты кодируется еще проще: все то же самое, только переменная носит другое имя, другая метка, отсутствуют значения, и выставляется количественная шкала (рис. 2.14).

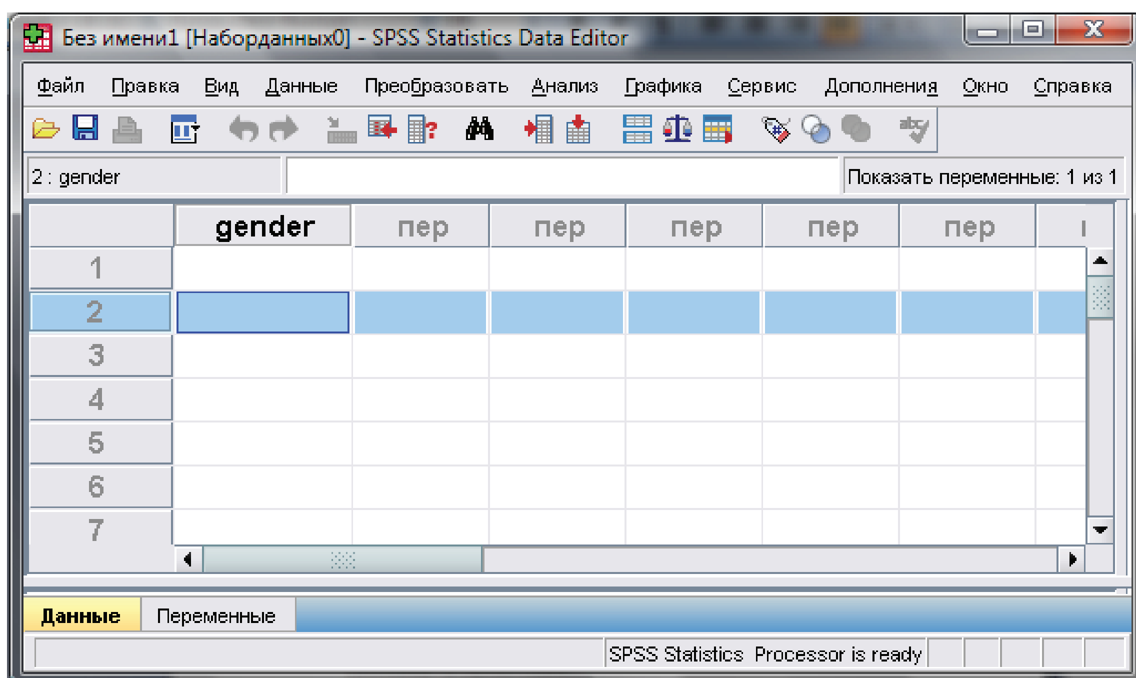


Рис. 2.13. Новая переменная в окне редактора данных

Файл Правка Вид Данные Преобразовать Анализ Графика Сервис Дополнения Окно Справка

	Имя	Тип	Ширина	Десятичные	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	gender	Числовая	1	0	Пол респондента	{1, Мужской}...	99	8	По право...	Номинальн...
2	age	Числовая	2	0	Возраст респондента	Нет	99	8	По право...	Количестве...
3										
4										
5										
6										
7										
8										
9										

Рис 2.14. Кодирование вопроса о возрасте респондента

Третий вопрос анкеты принципиально отличается от двух предыдущих. Он представляет собой вопрос с множественными ответами, то есть, как видно из формулировки, респондент может выбрать три варианта из предложенных ему ответов на вопрос. Соответственно, и кодируется он иначе, чем остальные. Особенность кодировки этого вопроса в том, что каждый вариант ответа представляет собой отдельную переменную (рис. 2.15).

Важным моментом является то, что значения к меткам в каждой переменной представлены всего тремя вариантами: 0 – «Не выбрано», 1 – «Выбрано» и 99 – «Нет ответа» (рис. 2.16).

Таким образом, эти метки несут в себе особую смысловую нагрузку, которая заключается в том, что за каждой меткой не стоит определенное значение, как в случае вопроса о поле респондента – метка играет роль своеобразной «галочки», указывающей на то, выбрал респондент тот или иной вариант ответа или вообще проигнорировал данный вопрос.

Файл Правка Вид Данные Преобразовать Анализ Графика Сервис Дополнения Окно Справка

	Имя	Тип	Ши...	...	Метка	Значения	П...	...	Выравнив...	Шкала
1	gender	Числовая	2	0	Пол респонде...	{1, Мужско...	99	8	По прав...	Номинальн...
2	age	Числовая	2	0	Возраст респ...	Нет	99	8	По прав...	Количестве...
3	a3_1	Числовая	2	0	Литературные...	{0, Не выб...	99	8	По прав...	Количестве...
4	a3_2	Числовая	2	0	Политики, гос...	{0, Не выб...	99	8	По прав...	Количестве...
5	a3_3	Числовая	2	0	Бизнесмены, ...	{0, Не выб...	99	8	По прав...	Количестве...
6	a3_4	Числовая	2	0	Преподавател...	{0, Не выб...	99	8	По прав...	Количестве...
7	a3_5	Числовая	2	0	Деятели куль...	{0, Не выб...	99	8	По прав...	Количестве...
8	a3_6	Числовая	2	0	Звезды шоу б...	{0, Не выб...	99	8	По прав...	Количестве...
9	a3_7	Числовая	2	0	Известные уч...	{0, Не выб...	99	8	По прав...	Количестве...
10	a3_8	Числовая	2	0	Светская "тус...	{0, Не выб...	99	8	По прав...	Количестве...
11	a3_9	Числовая	2	0	Звезды телев...	{0, Не выб...	99	8	По прав...	Количестве...
12	a3_10	Числовая	2	0	Родители, ро...	{0, Не выб...	99	8	По прав...	Количестве...
13	a3_11	Числовая	2	0	Друзья, знако...	{0, Не выб...	99	8	По прав...	Количестве...
14	a3_12	Текстовая	100	0	Свой вариант	{1, Часто}...	99	8	По лево...	Номинал...

Рис 2.15. Кодирование вопроса с множественными ответами

Метки значений

Метки значений

Значение:

Метка:

Проверка орфографии...

0 = "Не выбранно"

1 = "Выбранно"

99 = "Нет ответа"

Добавить

Изменить

Удалить

ОК Отмена Справка

Рис. 2.16. Кодирование вариантов ответов в вопросе с множественными ответами

Пропущенные значения, обозначенные цифрой «99», будут выставляться в окне Редактора данных только в том случае, если респондент не выбрал ни одного из вариантов.

Обратим внимание на то, что переменная a3_12 – «Другое (укажите свой вариант)» является открытым вопросом и предполагает внесение информации в текстовой форме, таким образом, в колонке «Тип переменной» при кодировке выбирается вариант «Текстовая», а в колонке «Ширина» указывается значение «100», которое ограничивает число вводимых символов до ста.

Следующий вопрос «Оцените, как часто на лекционных и семинарских занятиях в вашем вузе поднимаются следующие темы» представляет собой нечто среднее между вопросом с множественными ответами и закрытым вопросом с определенными вариантами ответов. Кодируется он также как и предыдущий рассматриваемый вопрос – каждый вариант как отдельная переменная, но в метках выставляются варианты ответов на этот вопрос. Таких в нашем случае пять: 1 – «Часто», 2 – «Иногда», 3 – «Поднимали 2–3 раза», 4 – «Практически не поднимаются», 5 – «Не задумывался над этим» и шестая, вспомогательная, метка 99 – «Нет ответа». Одновременно выставляется тип шкалы «Порядковая» (рис. 2.17).

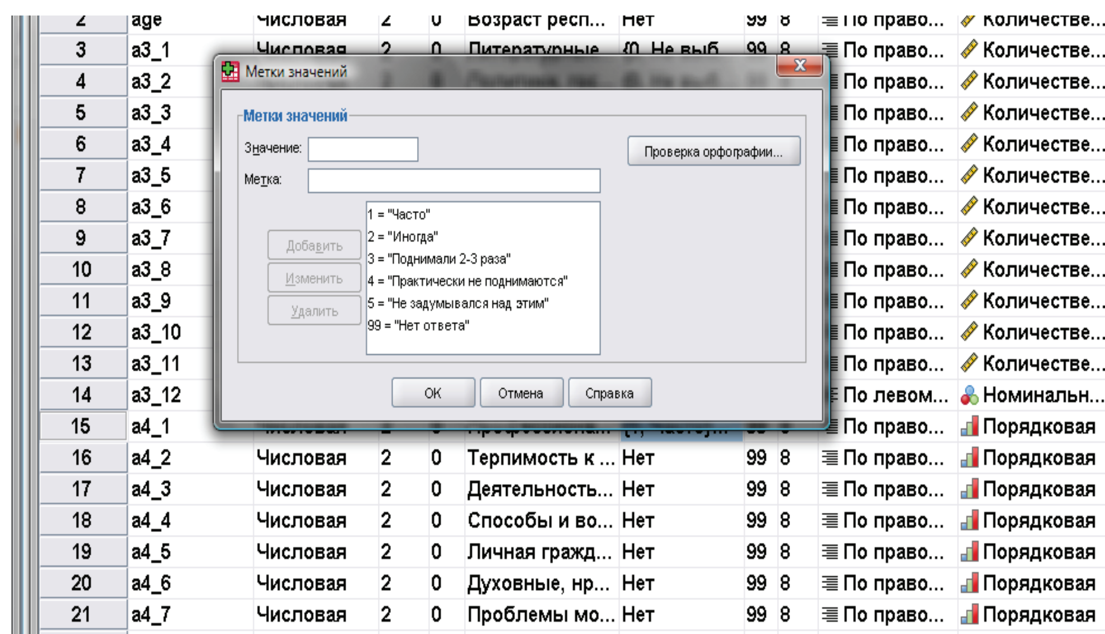
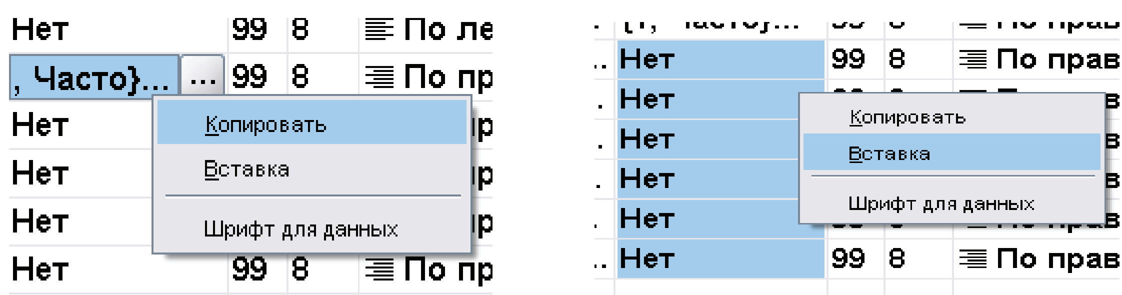


Рис. 2.17. Кодирование вопроса «Оцените, как часто на лекционных и семинарских занятиях в вашем вузе поднимаются следующие темы»

Для того, чтобы не вводить метки значений в каждую переменную, достаточно скопировать уже закодированные метки в соответствующей ячейке, выделить все ячейки, в которых будут использоваться те же метки, и вставить, используя правую кнопку мыши (рис. 2.18).



Копирование Вставка

Рис. 2.18. Копирование и вставка меток значений переменной

Последний вопрос нашей анкеты особо «замудреный». Кодировка этого вопроса требует определенного творчества и может быть закодирована несколькими различными способами. Ниже рассмотрим один из них. Для наглядности напомним содержание вопроса: «На какой ступеньке "сидят" сегодня большинство студентов и преподавателей в шкале "вечных" ценностей (обозначьте: С – студенты, П – преподаватели)»? Ниже представлена схема, с помощью которой респонденту предлагается ответить на данный вопрос.

Таблица 2.1

Вопрос анкеты о шкале «вечных» ценностей

1. Лестница честности	**2. Лестница справедливости**	**3. Лестница патриотизма**	**4. Лестница толерантности**

Итак, мы видим, что респонденту предлагается четыре лестницы, на каждой из которых необходимо отметить два варианта – студентов и преподавателей. Таким образом, в матрице SPSS будет кодироваться восемь переменных. Значения меток в этом случае отсутствуют (за исключением 99 – «Нет ответа»), а в окно Редактора данных будут вноситься численные значения от 1 до 10, что соответствует количеству ступеней в каждой лестнице (рис. 2.19). Тип шкалы для переменных в соответствующем столбце устанавливается «количественная».

Несмотря на то, что все вопросы анкеты в нашем примере уже закодированы, существует необходимость создания еще одной переменной, которая позволяет работать с данными более «комфортно». Это переменная, кодирующая номера анкет. Для чего это нужно? Достаточно часто в процессе статистической обработки данных возникает необходимость к возвращению к первичным опросным документам, т.е. к бумажным анкетам. Это происходит из-за того, что некоторые анкеты внесены в программу некорректно – либо в силу невнимательности людей, «забивающих» анкеты в программу, либо из-за неразборчивости почерка респондента или интервьюера, или по каким-то другим причинам. Кроме того, часто в процессе анализа обнаруживаются внутренние противоречия в ответах респондентов. В таких случаях исследователю необходимо найти бумажный вариант и выявить причину, сопоставив

данные анкеты и данные, внесенные в программу. Это практически невозможно осуществить, если перед началом обработки не пронумеровать весь массив (обычно это делается карандашом в углу титульной страницы анкеты) и не внести в SPSS каждую анкету под своим номером.

	Имя	Тип	...	...	Метка	Значения	П..
16	a4_2	Числовая	2	0	Терпимость к людям другой национ...	{1, Часто}...	99
17	a4_3	Числовая	2	0	Деятельность институтов гражданск...	{1, Часто}...	99
18	a4_4	Числовая	2	0	Способы и возможности самооргани...	{1, Часто}...	99
19	a4_5	Числовая	2	0	Личная гражданская ответственность	{1, Часто}...	99
20	a4_6	Числовая	2	0	Духовные, нравственные основы жи...	{1, Часто}...	99
21	a4_7	Числовая	2	0	Проблемы молодежной девиации (н...	{1, Часто}...	99
22	a5_1_1	Числовая	2	0	Лестница честности_студенты	{99, Нет ответ...	99
23	a5_1_2	Числовая	2	0	Лестница справедливости_студенты	{99, Нет ответ...	99
24	a5_1_3	Числовая	2	0	Лестница патриотизма_студенты	{99, Нет ответ...	99
25	a5_1_4	Числовая	2	0	Лестница толерантности_студенты	{99, Нет ответ...	99
26	a5_2_1	Числовая	2	0	Лестница честности_преподаватели	{99, Нет ответ...	99
27	a5_2_2	Числовая	2	0	Лестница справедливости_преподав...	{99, Нет ответ...	99
28	a5_2_3	Числовая	2	0	Лестница патриотизма_преподаватели	{99, Нет ответ...	99
29	a5_2_4	Числовая	2	0	Лестница толерантности_преподава...	{99, Нет ответ...	99
30							

Рис. 2.19. Кодирование вопроса о ценностях студентов и преподавателей

Обычно переменная, содержащая данные о номере анкеты, размещается первой. В нашем случае ее нужно вставить в уже существующую матрицу данных. Для этого правой кнопкой мыши нажимаем по переменной под номером 1 и в появившемся диалоговом окне выбираем пункт «Вставить переменную» (рис. 2.20).

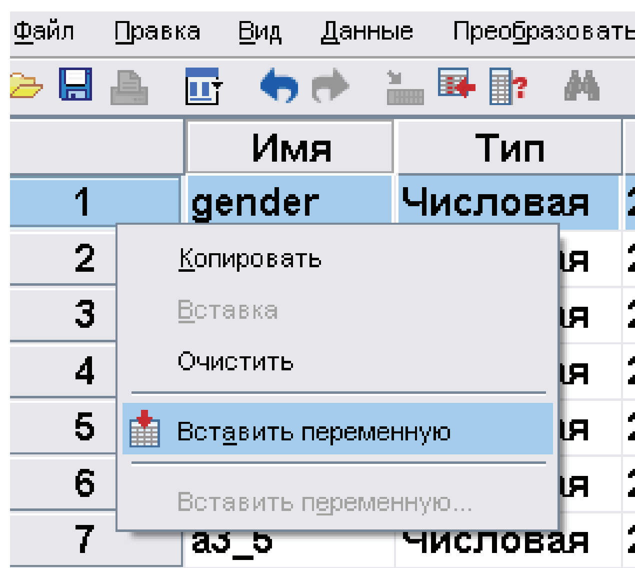


Рис. 2.20. Вставка переменной

В итоге появляется новая переменная, которую и кодируем для дальнейшего внесения в нее номеров анкет. Дадим ей имя «n», тип «числовая», ширина – 8 символов, метка – «Номер анкеты», выравнивание (для удобства) «по центру», шкала – «количественная» (рис 2.21).

йл Правка Вид Данные Преобразовать Анализ Графика Сервис Дополнения Окно Справка

	Имя	Тип	Шир...	Д...	Метка	Значения	Про...	Ст...	Выравнива...	Шкала
1	n	Числовая	8	0	Номер анкеты	Нет	Нет	8	По центру	Количе...
2	gender	Числовая	2	0	Пол респондента	{1, Мужск...	99	8	По право...	Номина...
3	age	Числовая	2	0	Возраст респонд...	Нет	99	8	По право...	Количе...
4	a3_1	Числовая	2	0	Литературные ге...	{0, Не вы...	99	8	По право...	Количе...

Рис. 2.21. Кодирование переменной «Номер анкеты».

Итак, матрица для нашей анкеты в SPSS создана и готова для внесения в нее результатов опроса. По данной анкете автором было проведено мини-исследование среди студентов одной группы специальности «Социология» Тюменского государственного нефтегазового университета с целью наглядного представления окна Редактора данных после внесения в него данных (рис. 2.22).

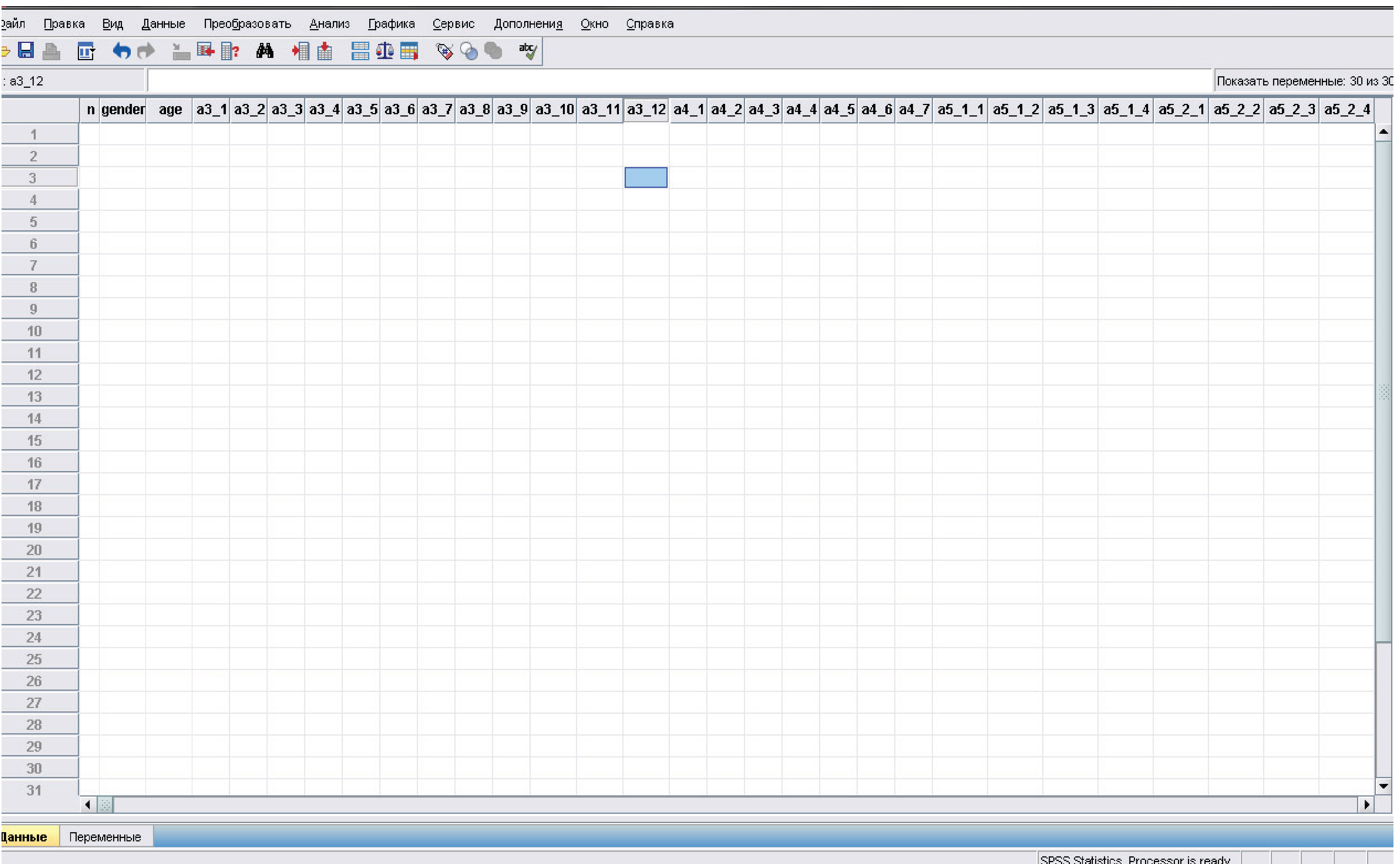


Рис. 2.22. Редактор данных с внесенным массивом данных, полученных в результате исследования

Вопросы и задания

1. Перечислите основные типы шкал, используемых в социологических исследованиях, и охарактеризуйте их.
2. Для чего в SPSS необходимо окно вывода?
3. На основании исследований, хранящихся в базе ЕАЭСД, создайте небольшую анкету по интересующей вас тематике, используя все типы шкал.
4. Закодируйте имеющуюся у Вас анкету в SPSS.
5. Если есть возможность, опросите несколько человек по вашей анкете и введите результаты в матрицу SPSS, которую Вы сами же и создали.

Список литературы

6. Бююль, А. SPSS: искусство обработки информации. Platinum Edition/А. Бююль, П. Цёфель. — М.: Изд-во «Диасофт», 2005. — 608 с.

7. Дубнов, П. Ю. Обработка статистической информации с помощью SPSS/П. Ю. Дубнов. — М.: Изд-во «НТ Пресс», 2004. — 221 с.

8. Калинин, С. И. Компьютерная обработка данных для психологов: Руководство/С. И. Калинин. — М.: Изд-во «Речь», 2002. — 136 с.

9. Крыштановский, А. О. Анализ социологических данных/А. О. Крыштановский. — М.: Изд-во «ГУ ВШЭ», 2007. — 281 с.

10. Наследов, А. Д. SPSS 15. Профессиональный статистический анализ данных/А. Наследов. — СПб: Изд-во «Питер», 2008. — 416 с.

11. Пациорковская, В. В. SPSS для социологов. Учебное пособие/В. В. Пациорковская, В. В. Пациорковский. — М.: Изд-во «ИСЭПН РАН», 2005. — 433 с.

12. Таганов, Д. Н. SPSS: статистический анализ в маркетинговых исследованиях/Д. Н. Таганов. — СПб.: Изд-во «Питер», 2005. — 192 с.

13. Турундаевский, В. Б. Многомерный статистический анализ в экономических задачах. Компьютерное моделирование в SPSS/В. Б. Турундаевский, И. В. Орлова, Н. А. Концевая. — М.: Изд-во «Вузовский учебник», 2009. — 320 с.

Лекция 3. Преобразование данных в SPSS

Преобразование данных является очень важной функцией в SPSS. В рамках данной функции доступно множество различных команд, основными из которых являются: вычисление переменных, подсчет значений в наблюдениях, перекодировка и категоризация переменных, ранжирование наблюдений, а также замена пропущенных значений (рис. 3.1). Рассмотрим некоторые из этих функции более подробно.

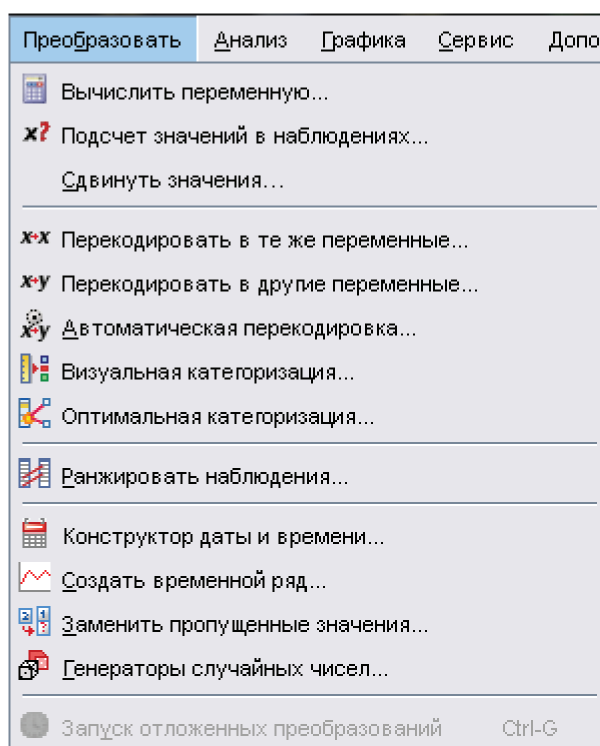


Рис. 3.1. Разделы меню «Преобразовать»

Вычисление переменных

Вычисление переменных позволяет существенно расширить возможности исследователя путем создания новых переменных или изменения значений существующих.

Ниже посмотрим, как осуществляются вычисления различных типов переменных и переменных с различными шкалами.

Вычисление количественных переменных. В лицензионных версиях программы SPSS имеется очень удобная встроенная система помощи пользователям, включающая в себя интерактивную справку и учебник. Хотелось бы привести пример именно из указанного учебника, т.к. он, по нашему мнению, очень наглядно демонстрирует процедуру вычисления количественных переменных.

Вычислять новые переменные можно, используя широкий спектр математических функций (в том числе даже очень сложные формулы). Для

примера мы произведем простое вычисление новой переменной, вычтя значения одной переменной из значений другой.

В файле данных demo.sav (идущим вместе с SPSS для обучающих целей) есть переменная «Возраст» (age) и переменная «Количество лет на текущем месте работы» (employ). А вот переменной, содержащей возраст во время поступления на последнее место работы, в этом файле нет. Мы можем создать новую переменную, в которой будет вычислена разность между возрастом в настоящее время и количеством лет на текущем месте работы, то есть примерный возраст в момент поступления на текущее место работы.

Для выполнения процедуры вычисления следуем алгоритму: Меню в окне Редактора данных → вкладка «Преобразовать» → пункт «Вычислить переменную» (рис. 3.2).

Рис. 3.2. Выбор раздела «Вычислить переменную» в меню «Преобразовать»

В появившемся диалоговом окне вводим в поле «Вычисляемая переменная» имя вычисляемой новой переменной «jobstart» (рис. 3.3).

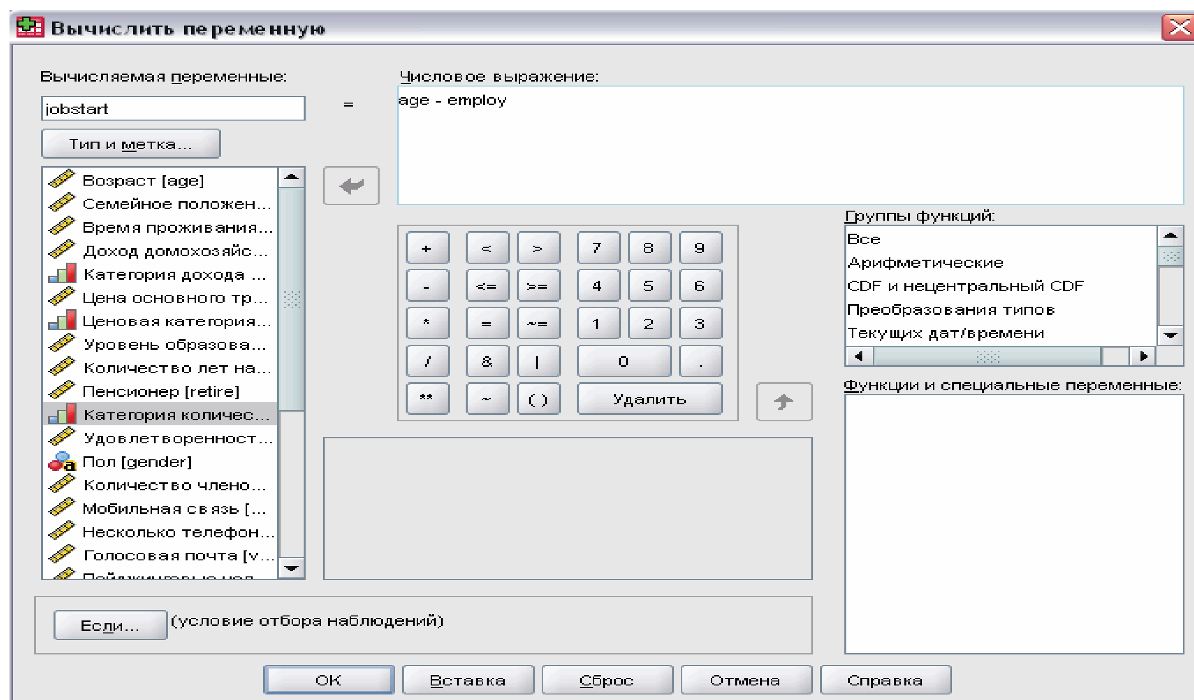


Рис. 3.3. Диалоговое окно «Вычислить переменную»

В списке исходных переменных выбираем «Возраст» (age) и копируем в поле «Числовое выражение» путем нажатия кнопки со стрелкой вправо. Далее нажимаем кнопку «минус» (–) на клавиатуре калькулятора в диалоговом окне (или клавишу «минус» на клавиатуре компьютера). Выбираем вторую переменную «Количество лет на текущем месте работы» (employ) и нажимаем кнопку в виде стрелки вправо, чтобы скопировать переменную в поле числового выражения.

Наконец, нажимаем «ОК», чтобы вычислить новую переменную.

В Редакторе данных появилась новая переменная. Поскольку новые переменные добавляются в конец файла, она находится в крайнем правом столбце в закладке «Данные» и в последней строке в закладке «Переменные» (рис. 3.4).

*demo.sav [Наборданных3] - Data Editor

Файл Правка Вид Данные Преобразовать Анализ Графика Сервис Окно Справка

1 : age | 55 | Показать переменные: 31 из 31

	ownfax	news	response	income2	jobstart
1	Нет	Да	Нет	50-74 тысяч	52,00
2	Нет	Да	Да	75+ тысяч	53,00
3	Нет	Нет	Нет	25-49 тысяч	27,00
4	Да	Нет	Нет	25-49 тысяч	23,00
5	Нет	Нет	Нет	До 25 тысяч	23,00
6	Нет	Да	Нет	75+ тысяч	43,00
7	Нет	Да	Нет	25-49 тысяч	40,00
8	Нет	Да	Нет	50-74 тысяч	34,00
9	Нет	Нет	Нет	До 25 тысяч	44,00
10	Да	Нет	Да	75+ тысяч	32,00
11	Нет	Нет	Нет	50-74 тысяч	54,00
12	Да	Нет	Нет	До 25 тысяч	27,00
13	Нет	Нет	Нет	25-49 тысяч	30,00
14	Нет	Да	Да	75+ тысяч	41,00

Данные Переменные

Рис. 3.4. Новая переменная в редакторе данных

Для числовых выражений можно также использовать встроенные функции. Функции разбиты на группы по своему назначению. К примеру, есть группа арифметических функций или группа статистических функций. В SPSS доступно около 70 встроенных функций, включая:

- арифметические функции;
- статистические функции;
- функции распределений;
- логические функции;
- функции агрегации и извлечения данных и времени;
- функции для работы с пропущенными значениями;

- функции для работы с несколькими наблюдениями;
- функции для работы с текстовыми значениями.

Для удобства в группы функций также включено несколько часто используемых системных переменных, таких как $TIME (текущая дата и время).

Краткое описание выбранной функции (в нашем случае SUM – суммирование) или системной переменной выводится в специальной области в диалоговом окне «Вычислить переменную» (рис. 3.5).

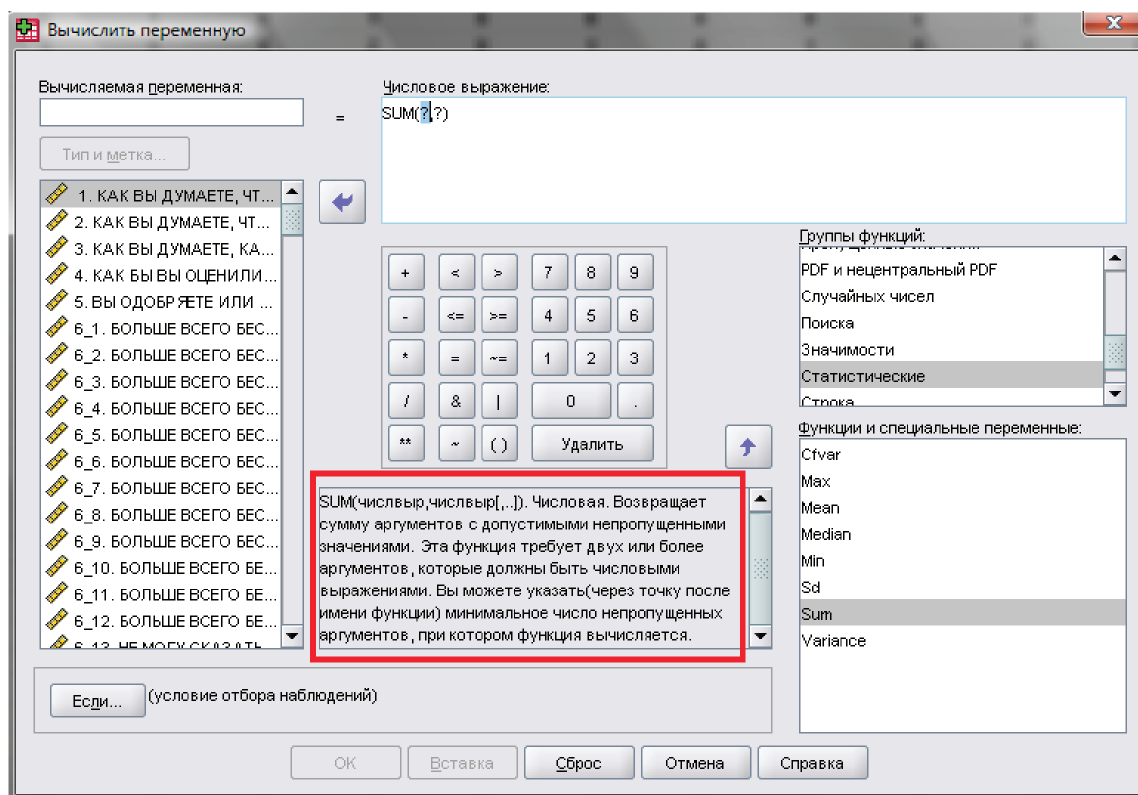


Рис. 3.5. Краткое описание выбранной функции

Для того чтобы вставить функцию в выражение, необходимо установить курсор в поле «Числовое выражение», далее выбрать в списке «Группы функций» подходящую группу. В группе «Все» представлены все доступные функции и системные переменные. В списке «Функции и специальные значения» нужно дважды щелкнуть по функции (или выбрать функцию и щелкнуть по стрелке рядом со списком «Группы функций»). Функция вставляется в выражение. Если выделит часть выражения, а затем вставить аргумент, выделенная часть выражения станет первым аргументом функции. Если аргументами являются имена переменных, их можно вставить из списка переменных.

Функция не является полной, пока не будут введены аргументы, представленные во вставленной функции знаками вопросов. Количество знаков вопроса указывает на минимальное количество аргументов, которые требуются, чтобы сделать функцию полной.

Задача условий для применения преобразований к подмножеству наблюдений. Для применения преобразований к подмножеству наблюдений используются условные выражения (также называемые логическими выражениями). Данная функция позволяет отбирать переменные, соответствующие определенным условиям.

Допустим, необходимо провести анализ ответов респондентов, которые старше 21 года. Для этого в диалоговом окне «Вычислить переменную» нажимаем на кнопку «Если», находящуюся в левом нижнем углу (рис. 3.6).

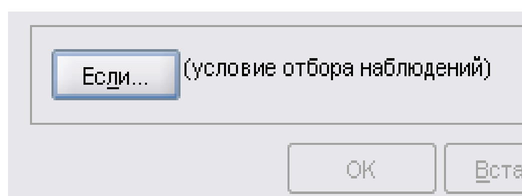


Рис. 3.6. Кнопка «Если» диалогового окна «Вычислить переменную»

В результате появляется диалоговое окно «Отбор наблюдений», схожее по внешнему виду с окном «Вычислить переменную» (рис. 3.7). В верхней части окна щелчком мыши активируем пункт «Включить наблюдения, удовлетворяющие условию», и в левой части из общего списка выбираем переменную «Возраст» (age), которую переносим в окно числового выражения.

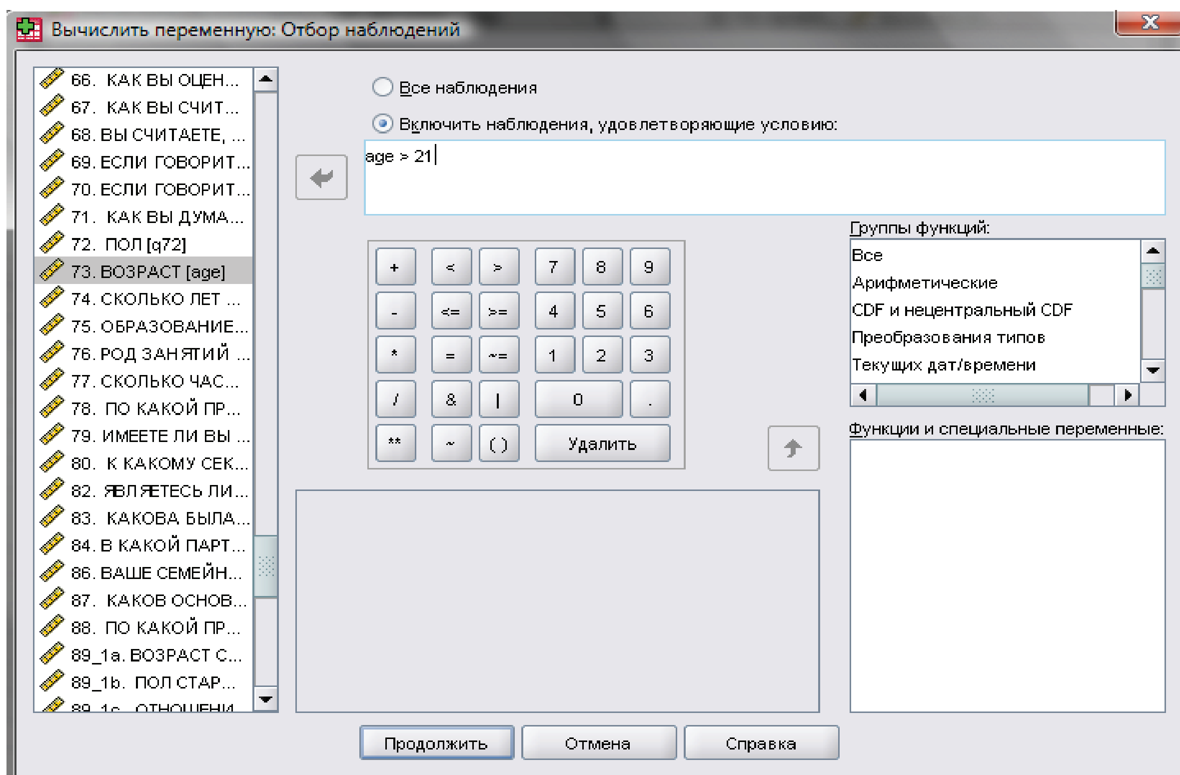


Рис. 3.7. Диалоговое окно «Вычислить переменную: Отбор наблюдений»

Далее при помощью клавиатуры калькулятора диалогового окна указываем требуемые условия – в нашем случае это «>21» – и нажимаем на кнопку «Продолжить».

Также можно связать несколько условных выражений, используя логические операторы, например, age>=21 | ed>=4 или income*3<100 & ed=5. В первом случае отбираются наблюдения, удовлетворяющие либо условию для переменной «Возраст» (age), либо условию для переменной «Уровень образования» (ed). Во втором примере, чтобы наблюдение было отобрано, должны выполняться оба условия: и для переменной «Доход домохозяйства в тысячах» (income), и для переменной «Уровень образования» (ed).

Аналогичным способом можно ввести любые условия, пользуясь всеми функциями процедуры «Вычислить переменную». В результате проделанных преобразований все последующие вычисления будут производиться с учетом заданных условий. То есть, в нашем примере, количество лет на текущем месте работы будет вычисляться только для тех респондентов, кто старше 21 года.

Вычисление переменных с порядковыми и номинальными шкалами. Существуют ситуации, когда необходимо произвести вычисление переменных, которые не являются количественными или текстовыми. Такие процедуры специфичны и требуют особых подходов. Рассмотрим на конкретном примере процедуру вычисления путем суммирования переменных, имеющих порядковую и номинальную шкалы.

Источником данных для примера выступила база данных социологического исследования, проведенного Всероссийским центром изучения общественного мнения в 1993 г. «Факт», методом самозаполнения по месту жительства по всероссийской многоступенчатой стратифицированной случайной выборке (объем выборочной совокупности – 1931) и было посвящено социальным проблемам, социальным тревогам и страхам. Указанная база находится в открытом доступе в едином архиве социологических данных СОФИСТ[8].

Переменными для процедуры суммирования выступили вопросы анкеты, посвященные религиозности респондентов, а именно два вопроса: «Считаете ли Вы себя религиозным человеком? Если да, то к какому вероисповеданию Вы себя относите?» и «Как часто Вы посещаете религиозные службы?».

Цель суммирования – построение шкалы, которая позволила бы выделить группы респондентов по степени религиозности. Для достижения поставленной цели необходимо предварительно провести ряд преобразований. Рассмотрим этот процесс более подробно.

---

8 http://sofist.socpol.ru/

Вопрос, направленный на определение религиозных убеждений респондента и включающий в себя несколько вариантов ответа (табл. 3.1), был преобразован в дихотомическую переменную, которая показывает только, верит респондент в Бога или не верит.

Таблица 3.1

Считаете ли вы себя религиозным человеком? Если да, то к какому вероисповеданию вы себя относите?

		Частота	Процент
Валидные	Не считаю себя религиозным человеком	773	40,0
	Отношу себя к православной вере	866	44,8
	Отношу себя к другой христианской вере	16	,8
	Отношу себя к иудаизму	4	,2
	Отношу себя к мусульманской вере	45	2,3
	Отношу себя к другой вере	11	,6
	Не могу указать вероисповедание	33	1,7
	Не хочу отвечать на этот вопрос	60	3,1
	Затрудняюсь ответить	123	6,4
	Всего	1931	100,0

Для этого варианты ответов, характеризующие принадлежность респондентов к той или иной религии и вариант «Не могу указать вероисповедание» были объединены в одну переменную – «Верующий». Вариант «Не считаю себя религиозным человеком» был перекодирован в «Не верующий», а варианты «Не хочу отвечать на этот вопрос» и «Затрудняюсь ответить» были отнесены к пропущенным значениям и, таким образом, исключены из анализа. В результате в матрице появилась новая переменная, линейное распределение которой, представлено в табл. 3.2.

Таблица 3.2

Отношение респондента к религии

		Частота	Процент
Валидные	Не верующий	773	40,0
	Верующий	975	50,5
	Всего	1748	90,5

Вторая переменная также подверглась процедуре перекодировки. Расположив варианты ответов в обратном порядке, а также исключив из вычислений такие варианты ответов, как: «Не хочу отвечать на этот вопрос» и «Затрудняюсь ответить» путем внесения их значений в «Пропуски», мы получили пригодную для суммирования переменную (табл. 3.3).

Таблица 3.3

Перекодированная частота посещаемости религиозных служб

		Частота	Процент
Валидные	Никогда	892	46,2
	Раз в году или реже	381	19,7
	Несколько раз в году	312	16,2
	Примерно раз в месяц	64	3,3
	Два-три раза в месяц	26	1,3
	Раз в неделю или чаще	14	,7
	Total	1689	87,5

Более подробно процедура перекодировки переменных будет рассмотрена ниже. Перед тем как произвести суммирование, посмотрим, как выглядит таблица сопряженности двух перекодированных переменных (табл. 3.4).

Таблица 3.4

Таблица сопряженности переменных «Частота посещения религиозных служб» и «Отношение респондента к религии»

Частота посещаемости религиозных служб		Отношение респондента к религии		Total
		Не верующий	Верующий	
Никогда	Count	575	240	815
	% по столбцу	82,5%	27,7%	52,1%
Раз в году или реже	Count	91	269	360
	% по столбцу	13,1%	31,0%	23,0%
Несколько раз в год	Count	30	257	287
	% по столбцу	4,3%	29,6%	18,4%
Примерно раз в месяц	Count	1	63	64
	% по столбцу	,1%	7,3%	4,1%
Два-три раза в месяц	Count	0	25	25
	% по столбцу	,0%	2,9%	1,6%
Раз в неделю или чаще	Count	0	13	13
	% по столбцу	,0%	1,5%	,8%
Total	Count	697	867	1564
	% по столбцу	100,0%	100,0%	100,0%

Тест на хи-квадрат Пирсона показал наличие взаимосвязи между двумя переменными. Коэффициент сопряженности же равен 0,489, что говорит о среднем уровне взаимосвязи между переменными.

Итак, перейдем к суммированию. Эта процедура позволит нам получить новую переменную, которая могла бы учитывать оба фактора: частоту посещения религиозных служб и веру в Бога. Для того чтобы процедура сум-

мирования имела смысл, необходимо присвоить коэффициент значимости для переменной «отношение респондента к религии». В нашем случае коэффициент 5 – для тех, кто верит в Бога, и 0 – для тех, кто не верит. Поясним, для чего это делается. Дело в том, что в процессе суммирования программа производит простое арифметическое действие, при этом не учитывая количество вариантов ответов в переменных с порядковой шкалой. Между тем это очень важный момент. В случае, если бы присвоение коэффициента «5» не было бы осуществлено, а варианты ответа остались бы под прежними номерами (0 – «Не верующий» и 1 – «Верующий»), то при сложении у нас получилось бы, с учетом нулевой группы, всего семь групп респондентов (табл. 3.5).

Таблица 3.5

Группировка верующих и неверующих респондентов в зависимости от посещаемости религиозных служб (до присвоения коэффициента)

	0. Не верующий	**1**. Верующий
0. Никогда	0 + 0 = 0 (0 группа)	1 + 0 = 1 (1 группа)
1. Раз в году или реже	0 + 1 = 1 (1 группа)	1 + 1 = 2 (2 группа)
2. Несколько раз в год	0 + 2 = 2 (2 группа)	1 + 2 = 3 (3 группа)
3. Примерно раз в месяц	0 + 3 = 3 (3 группа)	1 + 3 = 4 (4 группа)
4. Два-три раза в месяц	0 + 4 = 4 (4 группа)	1 + 4 = 5 (5 группа)
5. Раз в неделю или чаще	0 + 5 = 5 (5 группа)	1 + 5 = 6 (6 группа)

Процедура в таком случае потеряла бы всякий смысл, поскольку все данные хаотично перемешались. Таким образом, можно сделать ключевой вывод: при суммировании порядковых переменных необходимо учитывать количество вариантов ответов в каждой из них, и, в соответствии с этим, присваивать соответствующие коэффициенты. Присвоение коэффициентов – достаточно простая процедура, которая заключается в создании новой переменной, содержащей перекодированные варианты ответов.

Итак, присвоив коэффициент значимости 5 варианту «Верующий», при суммировании получается 11 групп респондентов (с учетом нулевой), которые с некоторой степенью условности можно охарактеризовать как различные по степени религиозности (табл. 3.6).

Из таблицы видно, что в нулевую группу будут входить респонденты, которые не верят в Бога и не посещают религиозных служб (наименее религиозная из групп). В пятую категорию входят респонденты, которые не верят в Бога, но при этом посещают религиозные службы; в эту же группу входят те, кто верит в Бога, но религиозные службы не посещает, их условно можно назвать «нейтральными». Наконец, в десятую группу входят те респонденты, которые верят в Бога и очень часто посещают религиозные службы – это наиболее религиозная группа.

Таблица 3.6

Группировка верующих и неверующих респондентов в зависимости от посещаемости религиозных служб (после присвоения коэффициента)

	0. Не верующий	**5**. Верующий
0. Никогда	0 + 0 = 0 (0 группа)	**5** + 0 = 5 (5 группа)
1. Раз в год или реже	0 + 1 = 1 (1 группа)	**5** +1 = 6 (6 группа)
2. Несколько раз в год	0 + 2 = 2 (2 группа)	**5** +2 = 7 (7 группа)
3. Примерно раз в месяц	0 + 3 = 3 (3 группа)	**5** +3 = 8 (8 группа)
4. Два-три раза в месяц	0 + 4 = 4 (4 группа)	**5** +4 = 9 (9 группа)
5. Раз в неделю или чаще	0 + 5 = 5 (5 группа)	**5** +5 = 10 (10 группа)

Линейное распределение новой переменной после осуществления процедуры суммирования в SPSS выглядит следующим образом (табл. 3.7):

Таблица 3.7

Сумма переменных, характеризующая индекс религиозности различных групп респондентов

		Частота	Процент
Валидные	0	575	29,8
	1	91	4,7
	2	30	1,6
	3	1	,1
	5	240	12,4
	6	269	13,9
	7	257	13,3
	8	63	3,3
	9	25	1,3
	10	13	,7
	Всего	1564	81,0
Системные пропущенные		367	19,0
Всего		1931	100,0

На этом процедура вычисления двух переменных завершена, однако для логического завершения необходимо осуществить еще одно действие. Проанализируем полученные данные. Как видно из таблицы, четвертая группа респондентов вообще выпадает из рассмотрения, т.к. среди опрошенных респондентов нет тех, кто не верит в Бога, но при этом два-три раза в месяц? посещает религиозные службы. Кроме того, в третью группу, характеризующуюся неверием в Бога, но ежемесячным посещением религиозных служб, входит всего один респондент. В принципе, полученное распределение вполне логично, однако это не совсем удобно для дальнейшего анализа. Исходя из этого, представляется целесообразным укрупнить полученные

группы. Логика подсказывает, что вполне возможно создать пять основных групп респондентов по степени религиозности (табл. 3.8).

В результате укрупнения групп мы добились двух позитивных эффектов: во-первых, появилась возможность более удобного анализа, во-вторых, за счет укрупнения произошло нивелирование групп по числу в них респондентов.

Таблица 3.8

Перегруппировка (укрупнение) исходных групп

Исходная группа	Характеристика респондентов исходной группы	Укрупненная группа	Характеристика укрупненной группы
0	Не верит и никогда не посещает служб	**0**	Не верит и никогда не посещает служб
1	Не верит, но посещает службы один раз в год или реже	**1**	Не верит, но время от времени посещает службы
2	Не верит, но посещает службы несколько раз в год		
3	Не верит, но посещает службы примерно раз в месяц		
5	Не верит, но посещает службы раз в неделю или чаще, и верит, но никогда не посещает службы	**2**	Не верит, но посещает службы раз в неделю или чаще; и верит, но посещает службы очень редко или не посещает совсем
6	Верит и посещает службы раз в году или реже		
7	Верит и посещает службы несколько раз в год	**3**	Верит и периодически посещает службы
8	Верит и посещает службы два-три раза в месяц		
9	Верит и посещает службы примерно раз в месяц	**4**	Верит и посещает службы часто
10	Верит и посещает службы раз в неделю или чаще		

Это хорошо видно в таблице линейного распределения новой переменной, появившейся в результате проделанной работы (табл. 3.9). Процедуру укрупнения можно произвести при помощи функции перекодировки переменных.

Отметим, что появившиеся в результате суммирования системные пропущенные значения, отображенные в таблице, являются результатом исключения из вычислений тех респондентов, которые проигнорировали или затруднились ответить на рассматриваемые вопросы.

Таблица 3.9

Категории респондентов по степени религиозности

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	0	575	29,8	36,8	36,8
	1	122	6,3	7,8	44,6
	2	509	26,4	32,5	77,1
	3	320	16,6	20,5	97,6
	4	38	2,0	2,4	100,0
	Total	1564	81,0	100,0	
Системные пропущенные		367	19,0		
Всего		1931	100,0		

Вычисления дихотомических переменных. С дихотомическими переменными все гораздо проще. Обратимся вновь к исследованию, которое использовалось для предыдущего примера. В частности, там есть такой вопрос:

Какие из внутренних проблем нашего общества беспокоят Вас больше всего? (дайте не более трех ответов)

1. Нехватка продуктов питания, товаров первой необходимости;
2. Рост цен;
3. Угроза безработицы;
4. Рост числа уголовных преступлений;
5. Кризис морали, культуры, нравственности;
6. Ухудшение состояния окружающей среды;
7. Обострение национальных конфликтов;
8. Уход от идеалов социализма;
9. Слабость, беспомощность государственной власти;
10. Угроза диктатуры;
11. Распад экономических связей между республиками бывшего Союза;
12. Другие (какие именно?);
13. Не могу сказать определенно.

Как видно из формулировки, вопрос представлен множественными вариантами ответов, т.к. респонденту можно выбрать три варианта. Таким образом, каждый из вариантов ответа в SPSS кодируется отдельной переменной, имеющей всего два варианта ответа (0 – «Не выбрано», 1 – «Выбрано»). Более подробно вопрос кодировки рассмотрен в Лекции 1.

Итак, допустим, что для решения определенной задачи нам необходимо выделить респондентов, которые выбрали первый и седьмой варианты

ответа одновременно. Проделав процедуру суммирования теперь уже известным нам способом, создается отдельная переменная (new). Линейное распределение данной переменной представлено в табл. 3.10.

Таблица 3.10

Переменная, полученная суммированием двух дихотомических переменных (new)

		Частота	Процент
Валидные	0,00	1315	68,1
	1,00	587	30,4
	2,00	29	1,5
	Итого	1931	100,0

Мы видим, что респондентов, которые не выбрали ни одного варианта ответа на вопрос, у нас больше всего – 1315 человек. или 68,1%. Те, кто выбрал один из двух рассматриваемых вариантов ответов – 30,4% и лишь 1,5% выбрали одновременно два варианта. Аналогичным образом мы можем сложить любое количество дихотомических переменных.

Подсчет встречаемости значений в наблюдениях

Данная функция позволяет создать новую переменную, в которой будет подсчитано количество одинаковых ответов в наборе переменных.

Для примера обратимся к базе данных исследования, проведенного ВЦИОМ в 1989 г. В этом исследовании респондентам предлагалось из списка, включающего в себя 22 наименования газет, выбрать те, которые он считает лучшими.

Для подсчета количества газет, которые каждый из респондентов считает лучшими, выбираем функцию «Подсчет значений в наблюдениях» в разделе меню «Преобразовать». В результате появляется диалоговое окно (рис. 3.8).

В окно «Вычисляемая переменная» вводим название создаваемой переменной (в нашем случае это «Newspaper»), в правом верхнем углу диалогового окна выставляем метку для этой переменной, и, наконец, при помощи стрелки выводим из общего списка переменных те, в которых необходимо подсчитать значения.

Далее нужно задать значения, которые будет считать программа. Делается это при помощи соответствующей кнопки «Задать значения», при нажатии на нее появляется диалоговое окно (рис. 3.9) . В нашем случае все переменные – дихотомические (варианты: 0 – «Не выбрано», 1 – «Выбрано»), поэтому указываем число 1 в окне «Подсчитываемые значения».

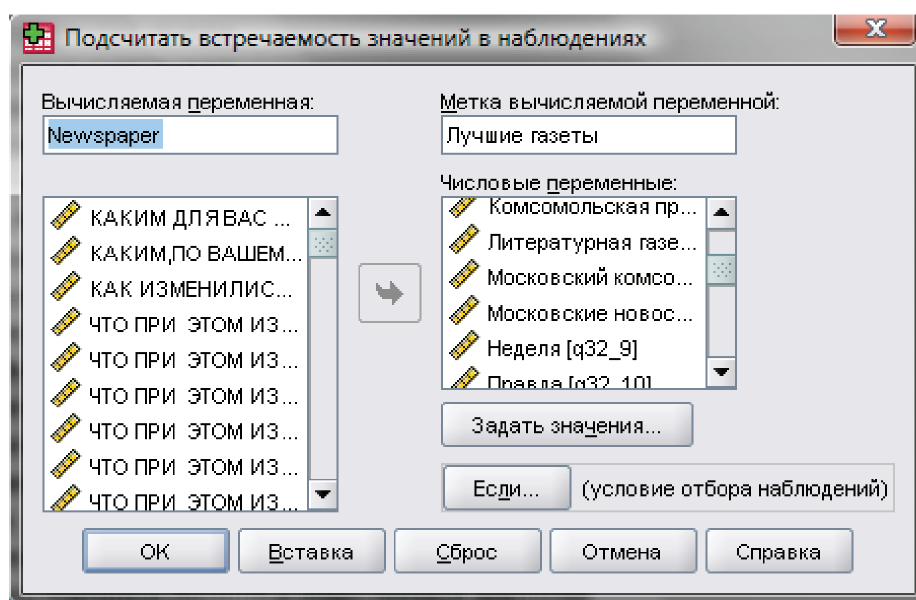


Рис. 3.8. Диалоговое окно «Подсчитать встречаемость в наблюдениях»

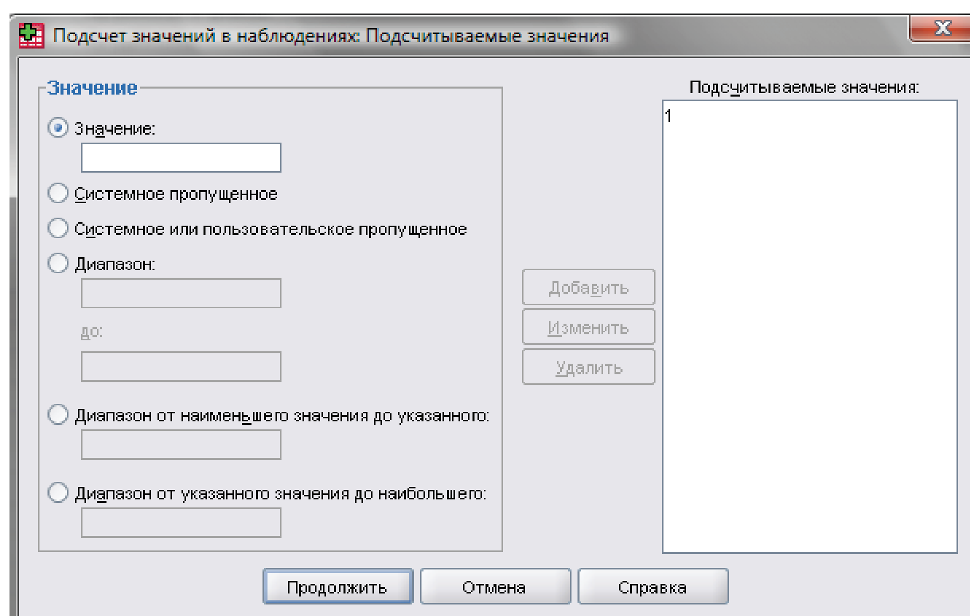


Рис. 3.9. Диалоговое окно «Подсчет значений в наблюдениях: Подсчитываемые значения»

Заданные значения могут быть отдельными значениями, пользовательскими или системными пропущенными значениями, а также интервалами (диапазонами). Диапазоны включают граничные значения и пользовательские пропущенные значения, попавшие в диапазон. Дополнительно в диалоговом окне «Подсчитать встречаемость значений в наблюдениях» есть возможность задать условие отбора наблюдений. Здесь мы не будем подробно останавливаться на этой функции, т.к. она была рассмотрена достаточно подробно в разделе о вычислении переменных.

Итак, в результате проделанных операций в редакторе переменных SPSS появляется новая переменная «Newspaper». Посмотрим, как выглядит частотное распределение данной переменной (табл. 3.11).

Таблица 3.11

Лучшие газеты

		Частота	Процент
Валидные	1,00	1940	94,4
	2,00	107	5,2
	3,00	6	,3
	4,00	1	,0
	Итого	2054	100,0

Из таблицы видно, что подавляющее большинство (94,4%) указало по одной газете, которую они считают лучшей. 5,2% указали по две такие газеты. Шестеро выбрали три, и один человек назвал четыре наименования газет, что в совокупности составило менее 1% от общего числа.

Необходимо отметить, что значение вычисляемой переменной, задаваемой в диалоговом окне «Подсчитать встречаемость значений в наблюдениях», увеличивается на единицу каждый раз, когда значение одной из переменных, заданных в списке «Числовые», совпадает с одним из заданных значений. Если в наблюдении встречается несколько переменных, значения которых совпадают с любым из заданных значений, вычисляемая переменная увеличивается на единицу столько раз, значения скольких переменных совпали с заданными значениями.

Перекодировка переменных

Данная функция позволяет перекодировать значения переменных или их диапазоны в новые значения. В SPSS существует три способа перекодировки: в те же или в другие переменные, а также автоматическая перекодировка. Рассмотрим реализацию данной функции на конкретном примере.

Выше мы описывали пример вычисления переменных, содержащих порядковые шкалы. В частности, говорилось о перекодировке вопроса, направленного на определение религиозных убеждений респондента. Вопрос, включающий в себя несколько вариантов ответа, нужно было преобразовать в дихотомическую переменную, которая показывает только то, верит респондент в Бога или не верит. Подробно эту процедуру мы не рассмотрели, сославшись на то, что это будет сделано ниже. Теперь же пришло время сделать это.

Итак, для осуществления перекодировки вызываем диалоговое окно, выбрав раздел «Перекодировать в другие переменные» в меню «Преобразовать». В разделе окна, находящемся слева из списка переменных, выбираем ту, которую необходимо перекодировать. В нашем случае это вопрос «Считаете ли Вы себя религиозным человеком? Если да, то к какому вероисповеданию

Вы себя относите?». При помощи стрелки переносим ее в центральную часть диалогового окна. Далее даем имя новой переменной (у нас – «religioznost»), задаем метку («Вера в Бога») и нажимаем кнопку «Изменить» (рис. 3.10).

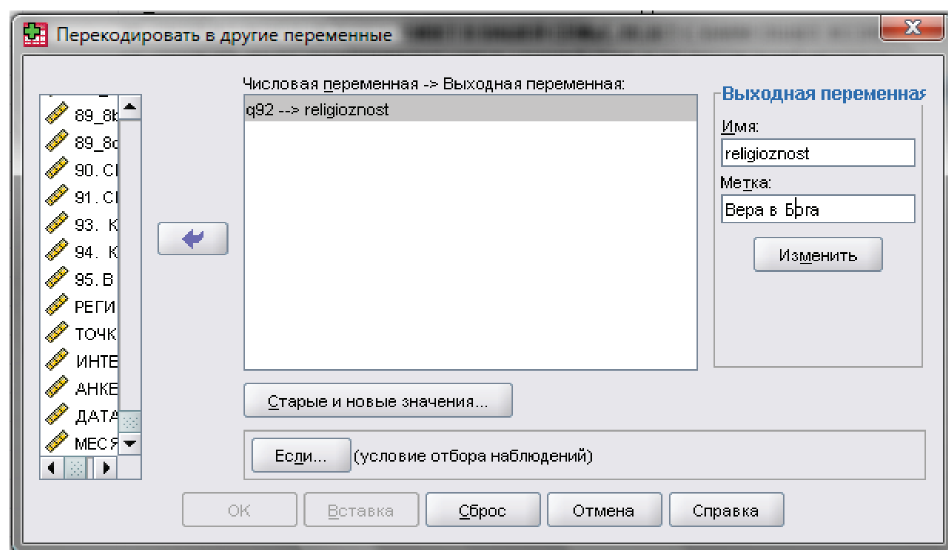


Рис. 3.10. Диалоговое окно «Перекодировать в другие переменные»

Теперь нужно заменить старые значения на новые. Для этого нажимаем кнопку «Старые и новые значения» в центральной нижней части окна. В результате появляется еще одно диалоговое окно, в котором и осуществляем описанную процедуру. Для того чтобы процедура перекодировки была более понятна, приведем старые метки значений переменной (рис. 3.11).

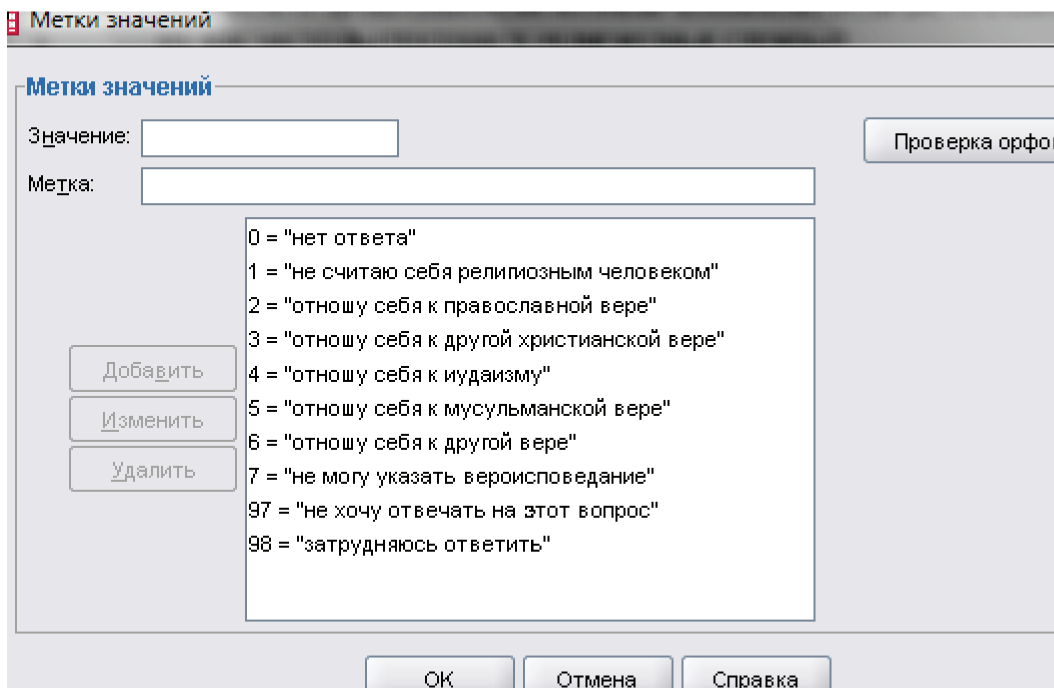


Рис. 3.11. Метки значений вопроса о религиозности респондентов

Варианты ответов, характеризующие принадлежность респондентов к той или иной религии (метки 2–6) и вариант 7 – «Не могу указать вероисповедание» нужно объединить в одну переменную – «Верующий». Вариант «Не считаю себя религиозным человеком» нужно перекодировать в «Не верующий», а варианты 0 – «Нет ответа», 97 – «Не хочу отвечать на этот вопрос» и 98 – «Затрудняюсь ответить» отнести к пропущенным значениям, и таким образом, исключить из анализа. Делается это очень просто – в левой части окна выбирается один из вариантов задаваемых значений и указывается само значение, в правой части указывается новое значение и нажимается кнопка «Добавить». Когда все изменения сделаны, нажимаем на кнопку «Продолжить» (рис. 3.12).

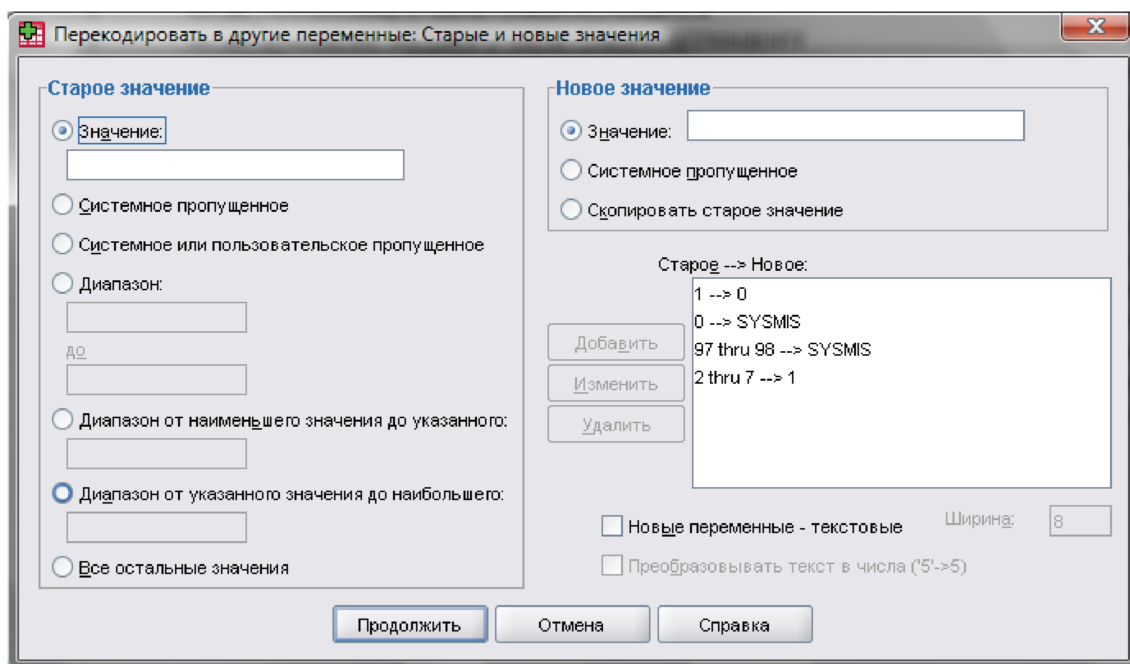


Рис. 3.12. Задача старых и новых значений в процессе перекодировки

Посмотрим, что же получилось. Появилась новая переменная «religioznost». В редакторе переменных зададим для нее следующие значения: 0 – «Не верю в Бога» и 1 – «Верю в Бога», и посмотрим частотное распределение (табл. 3.12).

Таблица 3.12

Вера в Бога

		Частота	Процент
Валидные	Не верю в Бога	773	40,0
	Верю в Бога	975	50,5
	Итого	1748	90,5
Пропущенные	Системные пропущенные	183	9,5
Итого		1931	100,0

В процессе перекодировки можно задавать определенные условия отбора наблюдений. Делается это уже рассмотренным нами способом, нажав кнопку «Если» в диалоговом окне «Перекодировать в другие переменные».

Перекодировку можно осуществлять сразу для нескольких переменных, однако должно соблюдаться важное условие – все они должны быть одного типа. Не допускается одновременное перекодирование числовых и текстовых переменных.

Как уже говорилось, в SPSS 17 существует три способа перекодировки. Один из них – «перекодировка в другие переменные» – мы рассмотрели. Не будем подробно останавливаться на двух остальных, рассмотрим лишь кратко их особенности.

Перекодировка в те же переменные. Особенность состоит в том, что при перекодировке не создается новая переменная, а изменяется старая. Процедура осуществляется почти таким же способом, как и та, которую мы описали выше.

Автоматическая перекодировка используется для преобразования текстовых и числовых значений в последовательные целые числа. Когда коды категорий переменной не являются последовательными, получившиеся пропущенные ячейки снижают производительность и увеличивают потребность в памяти при выполнении многих процедур SPSS. Кроме того, некоторые процедуры не могут использовать текстовые переменные, а некоторым процедурам непосредственно требуются последовательные целые числовые значения.

Категоризация переменных

Визуальная категоризация. Категоризация в какой-то мере тоже является перекодировкой, однако осуществляется по другому принципу. Категоризация необходима для создания новых переменных на основе группирования значений существующих переменных в ограниченное количество различающихся категорий. Эту процедуру можно использовать для создания новых переменных из непрерывных числовых переменных. Например, на основе количественной переменной «Возраст респондента» можно создать новую переменную, которая будет содержать удобное для исследователя количество возрастных категорий. Кроме того, процедура позволяет преобразовывать большое число категорий порядковой переменной в меньшее число категорий. Например, можно сократить оценку деятельности президента с десятибалльной шкалы до, например, трехбалльной: низкая, средняя и высокая.

Для примера используем данные исследования «Курьер» проведенного исследовательской организацией «Левада-центр» в 2007 г. (11-я волна). В частности, рассмотрим категоризацию переменной «Возраст респондента».

С целью осуществления необходимой процедуры вызываем стартовое диалоговое окно «Визуальное разбиение», выбрав пункт «Визуальная категоризация» в меню «Преобразовать» (рис. 1). Из общего списка выбираем переменную, которую необходимо категоризовать, при помощи кнопки со стрелкой переносим ее в соседнее окно «Переменные для категоризации» (рис. 3.13).

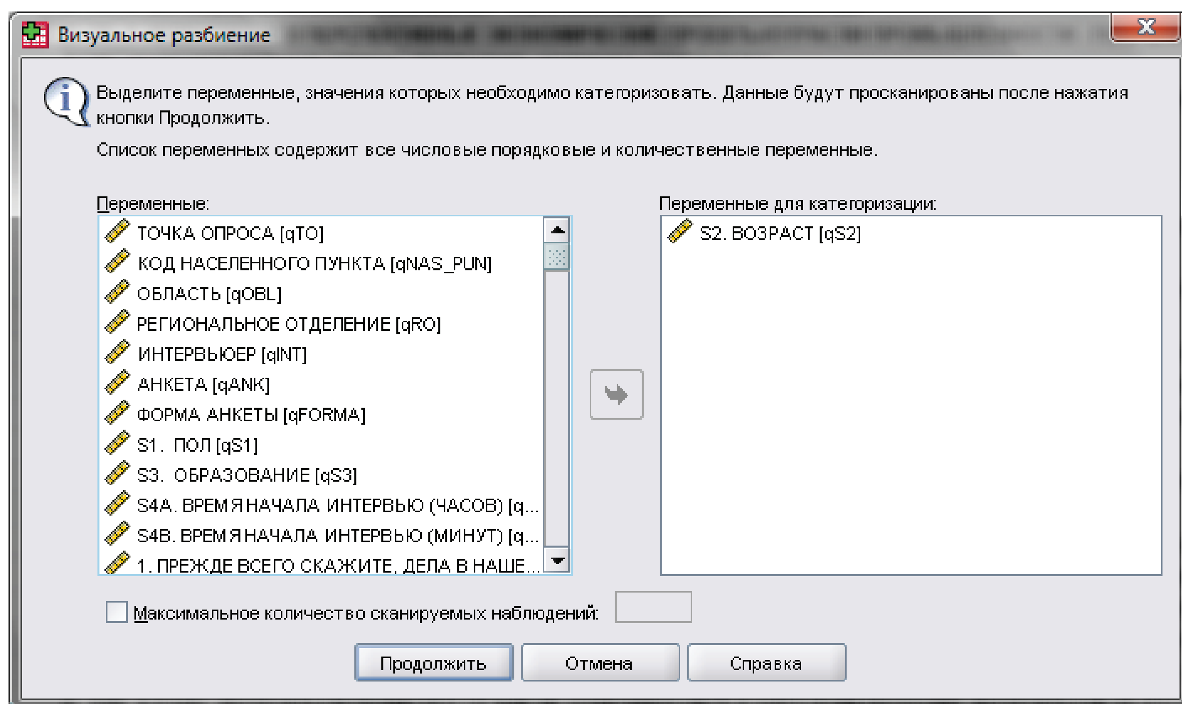


Рис. 3.13. Стартовое диалоговое окно «Визуальное разбиение»

Дополнительно в этом же окне можно ограничить число сканируемых наблюдений. В файлах данных с большим числом наблюдений ограничение числа сканируемых наблюдений может существенно сэкономить время, однако этого следует по возможности избегать, поскольку это влияет на распределение значений, используемых в последующих расчетах процедурой «Визуальное разбиение».

Еще одной особенностью является то, что текстовые переменные и номинальные числовые переменные не отображаются в списке исходных переменных. Процедура «Визуальное разбиение» требует числовых переменных, измеренных в количественной или порядковой шкале, поскольку предполагается, что значения данных имеют некоторый логический порядок, который можно использовать для естественной группировки значений. Если же все-таки возникает необходимость категоризации текстовых и номинальных переменных, то можно просто изменить тип шкалы нужной переменной в Редакторе переменных.

Итак, по завершении первого этапа нажимаем кнопку «Продолжить», после чего появляется основное диалоговое окно «Визуальное разбиение» (табл. 3.14).

Рис. 3.14. Основное диалоговое окно «Визуальное разбиение»

В этом окне представлено очень много информации. Попробуем разобраться в ней. Рассмотрим каждый раздел окна подробно.

1. Список отсканированных переменных. В списке выводятся переменные, которые мы выбрали в стартовом диалоговом окне. Их можно отсортировать по уровню измерений (количественный или порядковый), а также по имени или метке переменной, щелкнув по заголовку столбца.

2. Количество просканированных наблюдений и пропущенных значений. В этой части окна выводится число просканированных наблюдений и пропущенных значений. Все отсканированные наблюдения для выбранной переменной используются для формирования интервалов. Пропущенные же значения не включаются ни в одну из категорий интервалов.

2. Текущая и новая переменные. Здесь показывается имя и метка текущей выбранной переменной, а также имеется возможность введения имени и метки новой, разбиваемой переменной. По умолчанию для новой переменной используется метка или имя исходной переменной с добавленным словом «Категоризовано».

3. Минимум и максимум. Сразу под именами и метками переменных указываются минимальное и максимальное значения текущей выбранной

переменной по отсканированным наблюдениям, не включая пропущенные значения.

4. Гистограмма «Непропущенные значения». Этот график находится в центре диалогового окна и отображает распределение непропущенных значений текущей выбранной переменной на основе отсканированных наблюдений. После определения интервалов для новой переменной на гистограмме появляются вертикальные линии, обозначающие границы интервалов. В программе имеется возможность перетаскивать линии границ интервалов на гистограмме, изменяя ширину интервалов, а также удалять интервалы, перетаскивая линии границ за пределы гистограммы.

5. Сетка представлена таблицей, включающей три столбца. В первом столбце указан номер границы, во втором отображаются значения, определяющие верхние границы интервалов, в третьем – метки для каждого интервала. Ввести значения в эту таблицу можно вручную, или же воспользовавшись кнопкой «Границы интервалов», находящейся справа

Рассмотрим более подробно каждый из этих способов введения значений.

Ручной ввод. По умолчанию, автоматически включается граница интервала со значением «Высокий». Этот интервал будет включать все значения, превышающие значения остальных границ интервалов. Интервал, определяемый наименьшим значением границ интервалов, будет включать все значения: меньшие либо равные этому значению; просто меньше этого значения, – в зависимости от того, как определены верхние границы интервалов.

Вводим имя новой переменной «age». Установив курсор в первой строке второй колонки сетки, вводим при помощи клавиатуры числовое значение – допустим, «25». Это значение будет являться верхней границей первой категории, таким образом, сама категория будет включать все значения до 25. Учитывая, что в рассматриваемом исследовании опрос проводился среди респондентов старше 18 лет, первая категория будет включать в себя возрастную группу от 18 до 25 лет включительно. По мере того как мы вводим новое значение, программа автоматически добавляет новую строку в сетку. Во вторую строку вводим значение 37, в третью – 49, в четвертую – 58, и в пятую – 70. Одновременно мы наблюдаем появление на гистограмме вертикальных линий, которые позволяют визуально отслеживать процесс категоризации. Таким образом, у нас получилось шесть возрастных категорий. Метки значений, которые располагаются в третьей колонке, можно ввести вручную или воспользоваться кнопкой «Создать метки», чем мы и воспользовались (рис. 3.15).

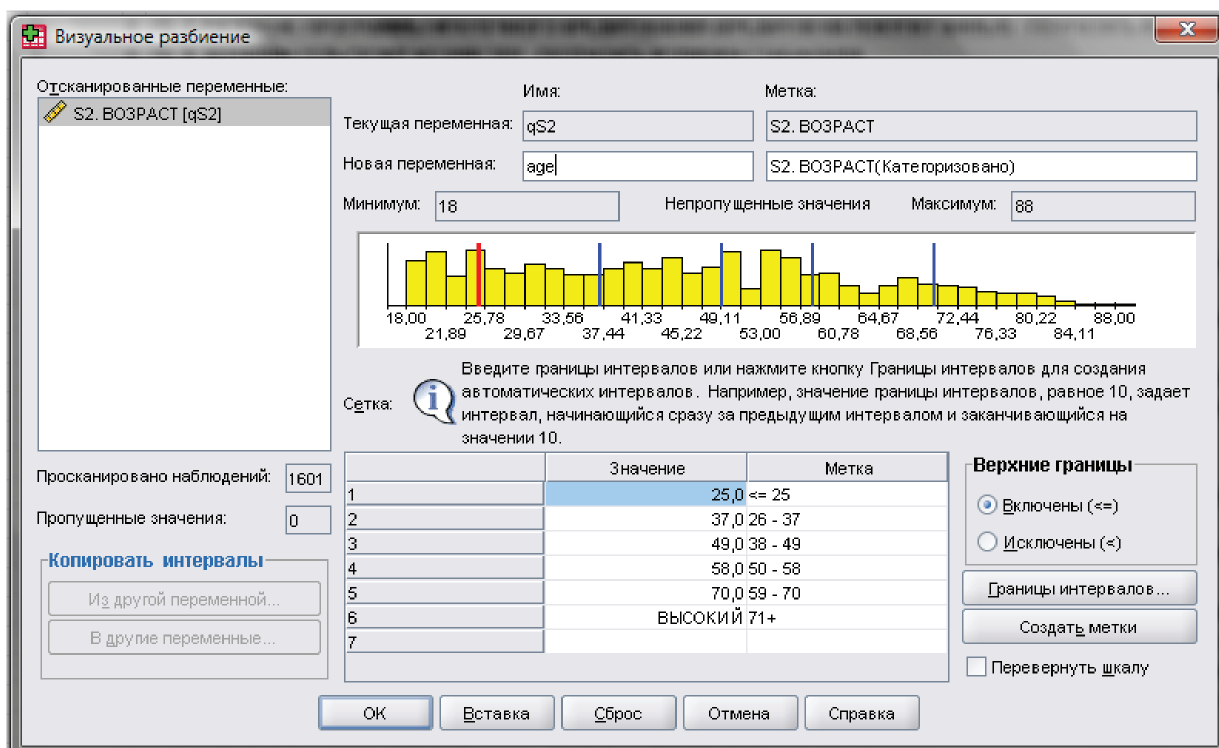


Рис. 3.15. Создание интервалов и меток «ручным» способом

Любой интервал или метка удаляется из сетки так же легко, как и создается. Для этого необходимо щелкнуть правой кнопкой мыши по ячейке «Значение» или «Метка» интервала, который необходимо удалить и из контекстного меню выбрать команду «Удалить строку». Отметим, что при удалении интервала «Высокий», всем наблюдениям, имеющим значения выше последнего значения границы интервала, в новой переменной будут назначены системные пропущенные значения. Выбрав команду «Удалить все метки» или «Удалить все границы», можно удалить все метки или все заданные интервалы.

В конечном итоге, при нажатии кнопки «ОК», программа создает новую переменную, которая является категоризованным возрастом респондентов.

Задача границ интервалов при помощи кнопки «Границы интервалов». При нажатии кнопки «Границы интервалов» появляется одноименное диалоговое окно (рис. 3.16).

Учитывая, что основной принцип осуществления процедуры категоризации нами усвоен на предыдущем примере, ниже просто опишем возможности автоматического формирования категорий интервалов при помощи диалогового окна «Границы интервалов» на основе выбранного критерия.

Раздел «Равные интервалы» формирует категории интервалов с равной шириной на основе любых двух из следующих трех критериев: «Местоположение первой границы» (значение, которое определяет верхнюю грани-

цу самого нижней категории интервала); «Количество границ»; «Ширина» каждого интервала (например, значение 10 разбило бы возраст в годах на интервалы по 10 лет).

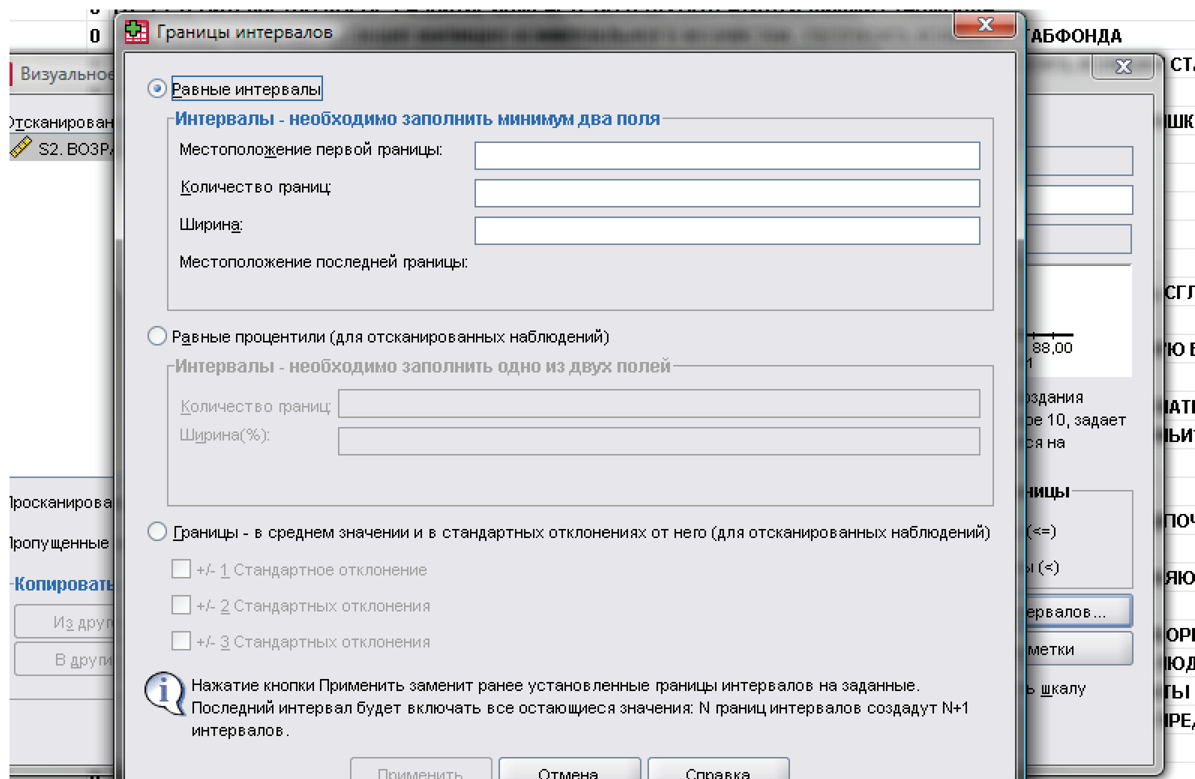


Рис. 3.16. Автоматическое формирование категорий интервалов

Раздел равные процентили (для отсканированных наблюдений) формирует категории интервалов с равным числом наблюдений в каждом интервале на основе одного двух критериев: «Количество границ»[9] и «Ширина»[10]. Если исходная переменная содержит относительно малое число различающихся значений или большое число одинаковых значений, может быть сформировано меньшее число интервалов, чем запрашивается. Если значению границы интервала соответствуют несколько идентичных значений, они все попадут в один и тот же интервал, поэтому фактический процент может не быть в точности равным тому, который должен выделяться.

Раздел «Границы – в среднем значении и в стандартных отклонениях от него (для отсканированных наблюдений)» формирует категории интер-

[9] Количество категорий интервалов равно количеству границ плюс единица. Например, три границы определяют четыре интервала процентилей (квартили), каждый из которых содержит по 25% наблюдений.

[10] Ширина каждого интервала, выражается в процентах от общего числа наблюдений. Например, значение 3,33 задавало бы три категории интервалов (две границы), каждый из которых содержал бы 33,3% наблюдений.

валов на основе значений среднего и стандартного отклонения распределения значений переменной. Можно выбрать любую комбинацию интервалов на основе одного, двух и/или трех стандартных отклонений. При этом если не выбран ни один из предложенных интервалов стандартных отклонений, формируются две категории интервалов с границей в среднем значении переменной.

Отметим, что расчеты процентилей и стандартных отклонений выполняются на основе отсканированных наблюдений. Если на первом этапе (отбора переменных для категоризации) число сканируемых наблюдений было ограниченно, то результирующие интервалы могут не содержать точную долю наблюдений, которую хотелось бы видеть в интервалах, в особенности, если файл данных был отсортирован по исходной переменной. Например, если мы ограничили число сканируемых наблюдений первыми 100 наблюдениями в файле данных, содержащем 1000 наблюдений, который был отсортирован в порядке возрастания значений возраста респондента, то вместо четырех интервалов, каждый из которых содержит по 25% наблюдений, мы можем обнаружить, что первые три интервала содержат примерно по 3,3% наблюдений каждый, тогда как последний содержит 90% наблюдений.

Теперь вернемся к остальным возможностям, предлагаемым SPSS, для удобства визуальной категоризации.

6. Перевернуть шкалу. По умолчанию, значения новой, категоризованной переменной являются последовательными целыми числами от 1 до n. Переворот шкалы приводит к последовательности целых чисел от n до 1.

7. Копировать интервалы. Данная функция позволяет копировать спецификации интервалов из другой переменной в текущую выбранную переменную или из выбранной переменной в несколько других переменных.

Оптимальная категоризация представляет собой особую процедуру категоризации одной или нескольких переменных путем распределения значений переменных в блоки. Например, переменная «уровень образования» является оптимальной по отношению к категориям переменной «должность». Блоки могут быть использованы вместо первоначальных значений данных для дальнейшего анализа в процедурах, которые требуют категориальных переменных. В рамках настоящей работы не будем подробно рассматривать данную функцию.

В заключение лекции отметим, что, наряду с оптимальной категоризацией, еще несколько функций в настоящем издании остались без внимания. Это связанно с тем, что, исходя из опыта авторов, в практике анализа данных социологических исследований эти функции используются достаточно редко. Соответственно, и приоритетность их по сравнению с рассмотренными

функциями гораздо ниже. Стоит добавить, что вопросы, не рассмотренные в этом разделе подробно, изучить самостоятельно не составит труда.

Вопросы и задания

1. Какими функциями по преобразованию данных обладает SPSS? Перечислите и кратко охарактеризуйте их.
2. В чем принципиальное отличие вычисления переменных с различными типами шкал?
3. В каких случаях исследователь использует функцию подсчета встречаемости в наблюдениях?
4. Для чего нужна категоризация переменных? Приведите примеры, когда в социологических исследованиях может пригодиться категоризация переменных.

Список литературы

1. Бююль, А. SPSS: искусство обработки информации. Platinum Edition/А. Бююль, П. Цёфель. — М.: Изд-во «Диасофт», 2005. — 608 с.

2. Дубнов, П. Ю. Обработка статистической информации с помощью SPSS/П. Ю. Дубнов. — М.: Изд-во «НТ Пресс», 2004. — 221 с.

3. Калинин, С. И. Компьютерная обработка данных для психологов: Руководство/С. И. Калинин. — М.: Изд-во «Речь», 2002. — 136 с.

4. Крыштановский, А. О. Анализ социологических данных/А. О. Крыштановский — М.: Изд-во «ГУ ВШЭ», 2007. — 281 с.

5. Наследов, А. Д. SPSS 15. Профессиональный статистический анализ данных/А. Наследов. — СПб: Изд-во «Питер», 2008. — 416 с.

6. Пациорковская, В. В. SPSS для социологов. Учебное пособие/В. В. Пациорковская, В. В. Пациорковский. — М.: Изд-во «ИСЭПН РАН», 2005. — 433 с.

7. Таганов Д. Н. SPSS: статистический анализ в маркетинговых исследованиях/Д. Н. Таганов. — СПб.: Изд-во «Питер», 2005. — 192 с.

8. Турундаевский, В. Б. Многомерный статистический анализ в экономических задачах. Компьютерное моделирование в SPSS/В. Б. Турундаевский, И. В. Орлова, Н. А. Концевая. — М.: Изд-во «Вузовский учебник», 2009. — 320 с.

Лекция 4. Частотный анализ

Собрав и введя в компьютер массив первичных социологических материалов, исследователь вплотную подходит к необходимости получения обобщенной информации о собранных данных. Для того чтобы получить самую общую картину по результатам проведенного исследования, социолог прибегает к простому частотному анализу. Этот анализ – лишь первый шаг на сложном пути изучения общества и социальных групп, их особенностей. Тем не менее, это очень важный этап в обработке первичных социологических данных, ведь именно благодаря анализу простых частотных распределений социолог может сориентироваться в собранном материале, наметить дальнейшие пути его анализа.

Пакет программ SPSS открывает широкие возможности не только для изучения частотных распределений, но и для их представления в табличной и графической формах, а также для проведения статистических тестов, которые позволяют получить дополнительные характеристики рассматриваемых переменных.

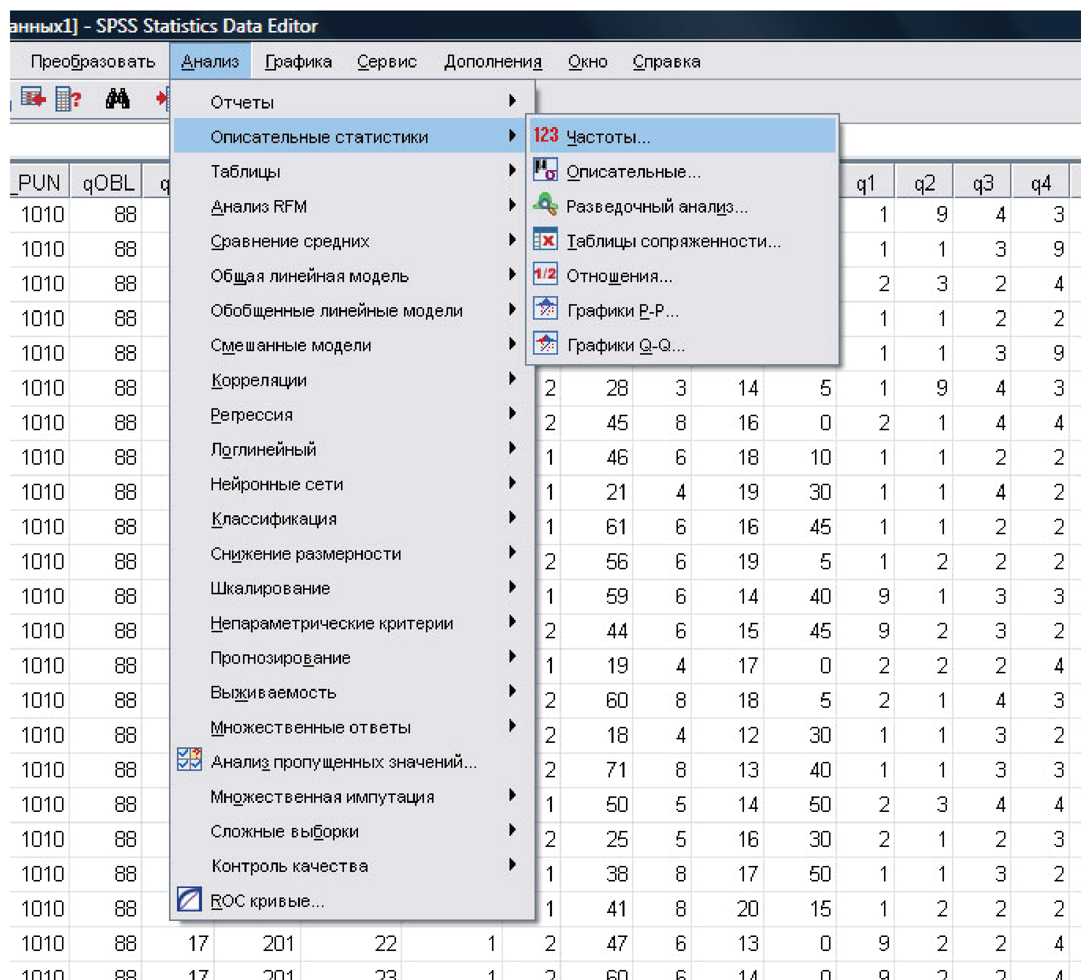


Рис. 4.1. Построение одномерных частотных таблиц с помощью меню «Анализ: Описательные статистики: Частоты»

Для проведения частотного анализа в SPSS необходимо воспользоваться командой «Частоты». В меню «Анализ» выберем пункт «Описательные статистики», а в ней – «Частоты» (см. рис. 4.1). После этого на мониторе вашего компьютера вы увидите следующее диалоговое окно (рис. 4.2), с которым нам и предстоит работать. В левой части окна расположен список доступных для анализа переменных, правая его часть пока пуста. В него с помощью стрелки мы перенесем ту переменную (или переменные), частотное распределение которой (или которых) мы хотим увидеть. В нашем случае это будет переменная q2 с результатами ответов на вопрос «Не могли бы Вы сказать, Вы, Ваша семья уже приспособились к переменам, произошедшим в стране в течение последних 10 лет; или думаете, что приспособитесь в ближайшем будущем; или думаете, что так никогда и не сможете к ним приспособиться?».

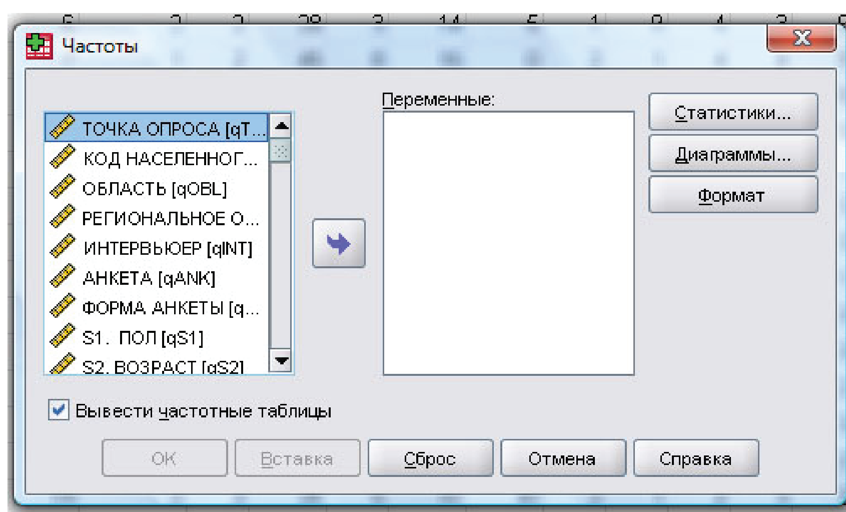


Рис. 4.2. Диалоговое окно «Частоты»

Выбрав эту переменную и нажав «ОК», получим одномерные таблицы 4.1 и 4.2. Из первой таблицы следует, что валидных значений 1601, пропущенных – ни одного, то есть ответы всех респондентов учтены в обработке.

Таблица 4.1

Статистики процедуры «Частоты»

Статистики		
2. НЕ МОГЛИ БЫ ВЫ СКАЗАТЬ, ВЫ, ВАША СЕМЬЯ УЖЕ ПРИСПОСОБИЛИСЬ К ПЕРЕМЕНАМ, ПРОИЗОШЕДШИМ В СТРАНЕ В ТЕЧЕНИЕ ПОСЛЕДНИХ 10 ЛЕТ; ИЛИ ДУМАЕТЕ, ЧТО ПРИСПОСОБИТЕСЬ В БЛИЖАЙШЕМ БУДУЩЕМ; ИЛИ ДУМАЕТЕ, ЧТО ТАК НИКОГДА И НЕ СМОЖЕТЕ К НИМ ПРИСПОСОБИТЬСЯ?		
N	Валидные	1601
	Пропущенные	0

Таблица 4.2

Распределение отчетов респондентов на вопрос: «Не могли бы Вы сказать, Вы, Ваша семья уже приспособились к переменам, произошедшим в стране в течение последних 10 лет; или думаете, что приспособитесь в ближайшем будущем; или думаете, что так никогда и не сможете к ним приспособиться?»

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	приспособились	978	61,1	61,1	61,1
	приспособимся в ближайшем будущем	226	14,1	14,1	75,2
	никогда не сможем приспособиться	313	19,6	19,6	94,8
	затрудняюсь ответить	84	5,2	5,2	100,0
	Итого	1601	100,0	100,0	

Данные таблицы 4.2 показывают, как именно отвечали респонденты на поставленный вопрос, какие варианты ответов предпочитали с какой частотой (как в абсолютном, так и в относительном исчислении). Из нее мы видим, что 978 человек приспособились к переменам, произошедшим в стране за последние 10 лет, чего не скажешь о 226 респондентах, которые считают, что приспособятся в ближайшем будущем, 313 респондентах, которые уверены, что вообще не приспособятся и 84-х, затруднившихся ответить.

Однако данные, представленные в столбце «Частота» не позволяют оценить масштабы предпочтений респондентов, если их не соотносить с общим числом опрошенных. Делать это постоянно неудобно, поэтому логичнее воспользоваться данными столбца «Процент», где представлены относительные данные о выборах респондентов. Отсюда мы видим, что большинство, или 61,1%, уже приспособились к переменам последних лет, 14,1% собираются приспособиться в ближайшем будущем, 19,6% никогда не приспособятся. Из этих данных уже складывается более понятная картина. Она может быть дополнена данными кумулятивных (накопленных) частот, представленных в последнем столбце. Кумулятивные проценты позволяют быстро оценить масштабы распространения явления среди нескольких групп респондентов. Так, мы видим, что 75,2% опрошенных уже приспособились или собираются приспособиться к произошедшим переменам. Наличие кумулятивных данных очень удобно для социолога, так как позволяет ему сократить затраты времени на подобные вычисления вручную.

Еще один важный столбец с данными – «Валидный процент». В нем представлены данные о значимом процентном соотношении разных групп респондентов без учета неответивших. В нашем случае таких не было, однако социолог в исследовательской практике часто сталкивается с обратной ситуацией.

Простые частотные данные можно представить и в графическом виде, причем варианты последнего могут варьироваться в зависимости от задач исследователя и его предпочтений. Графическое представление частотных распределений особенно важно, когда социологу необходимо представить результаты исследования для широкой, как правило, неподготовленной, аудитории, так как в таком виде они легче воспринимаются. Сразу покажем это на примере. Для того чтобы графически представить результаты частотного распределения воспользуемся Конструктором диаграмм из меню «Графика» (рис. 4.3 и 4.4).

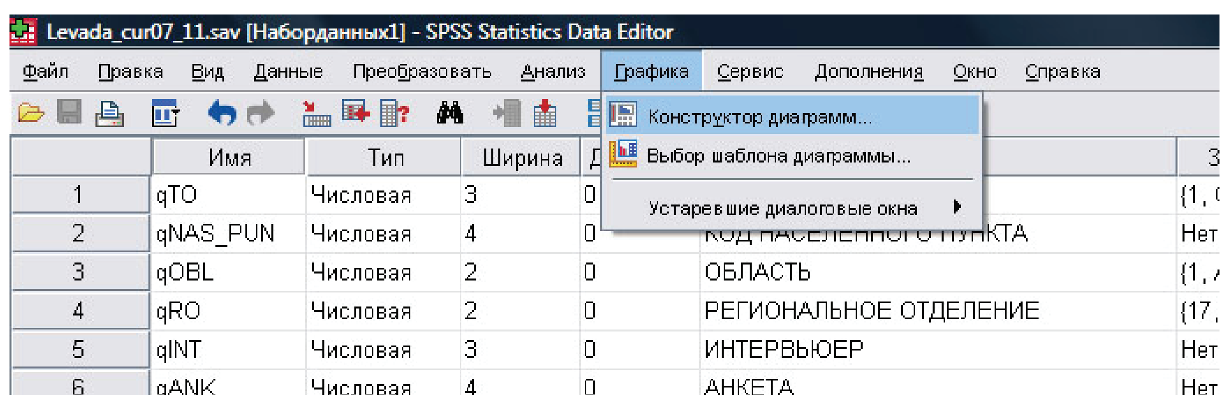


Рис. 4.3. Запуск конструктора диаграмм из меню «Графика»

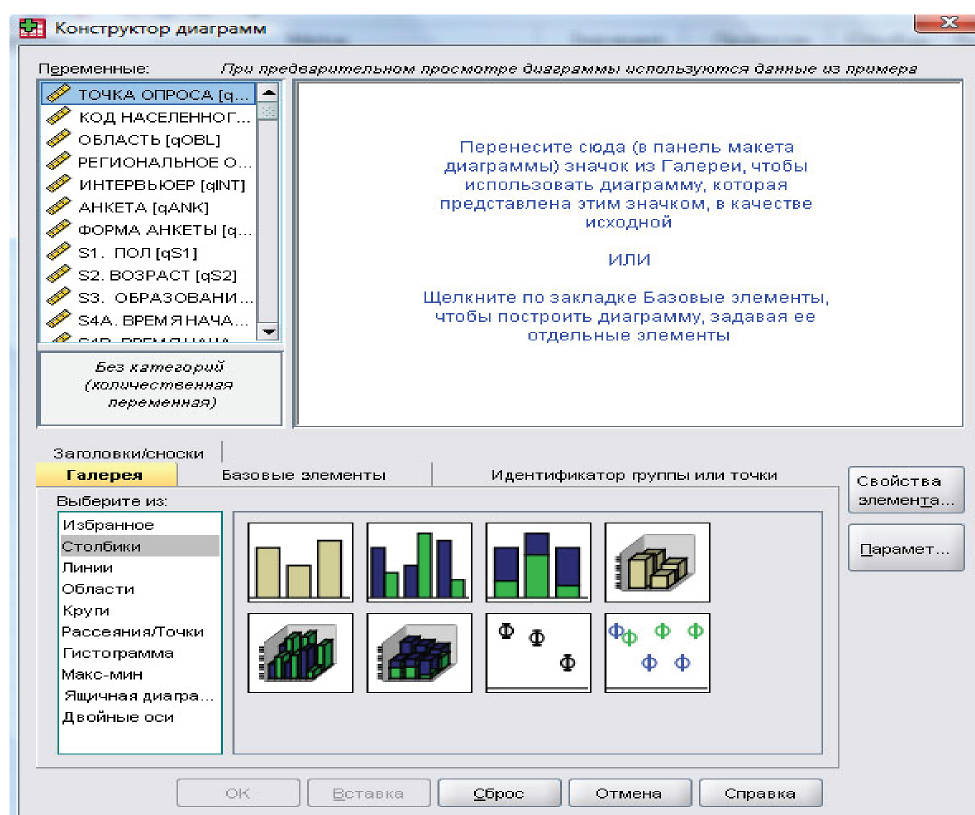


Рис. 4.4. Конструктор диаграмм

Конструктор диаграмм позволяет создавать диаграммы и графики в интерактивном режиме. Покажем возможности Конструктора диаграмм на нашем примере.

Для начала выберем в «Галерее» (нижняя часть конструктора) тип графика «Столбики» и перенесем мышью соответствующий значок в центральное поле конструктора. Конструктор диаграмм примет следующий вид (рис. 4.5).

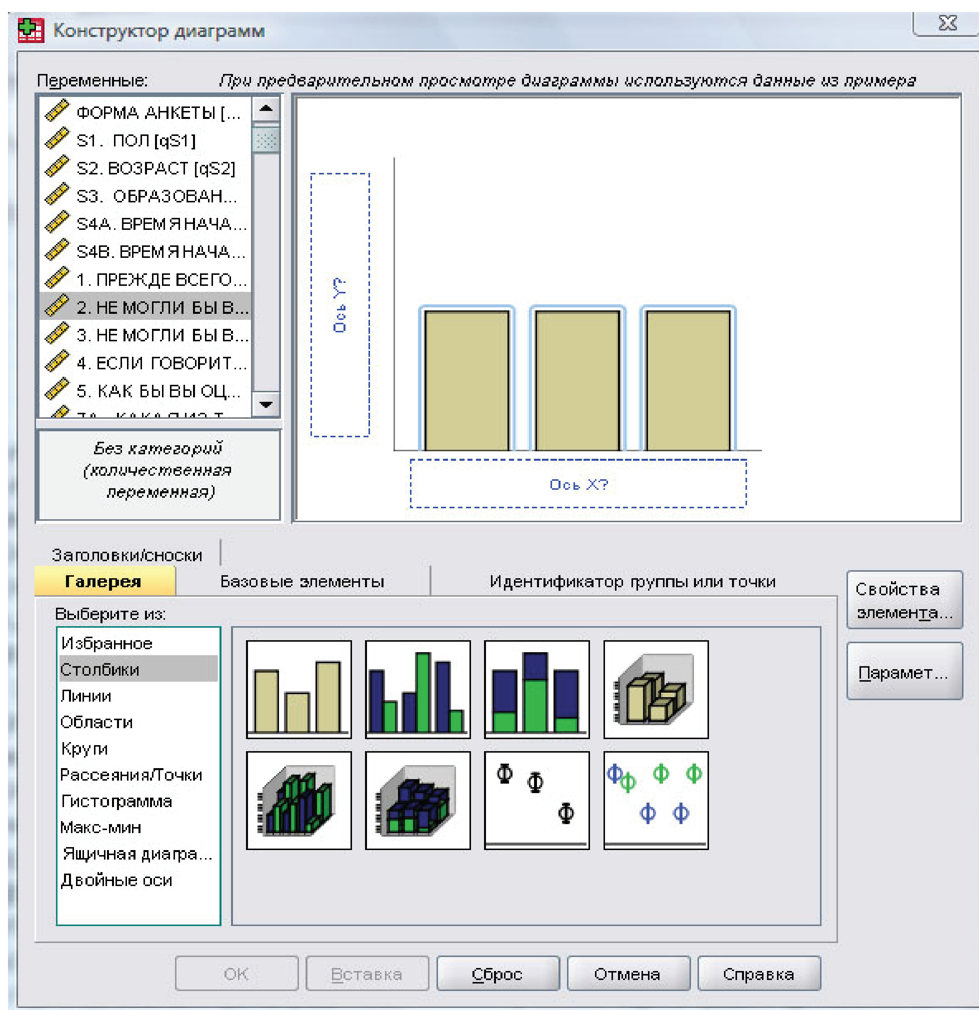


Рис. 4.5. Конструктор диаграмм после выбора вида диаграммы

Рядом с окном Конструктора диаграмм открылось еще одно диалоговое окно программы, которое поможет нам достроить график (рис. 6).

Но пока займемся выбором и подготовкой переменной для построения диаграммы. Найдем нашу переменную q2 в списке переменных в окне конструктора (рис. 4.5). Мы видим, что здесь она определена как количественная переменная, однако мы знаем, что наша переменная – номинальная. Выделив ее и нажав правую кнопку мыши, выберем пункт «Номинальная переменная» в контекстном меню. Это необходимо проделать, чтобы правильно построить график. Теперь под окном переменных в конструкторе диаграмм

появились варианты ответов на данный вопрос анкеты. Теперь переместим переменную q2 на ось X в центральном поле конструктора диаграмм.

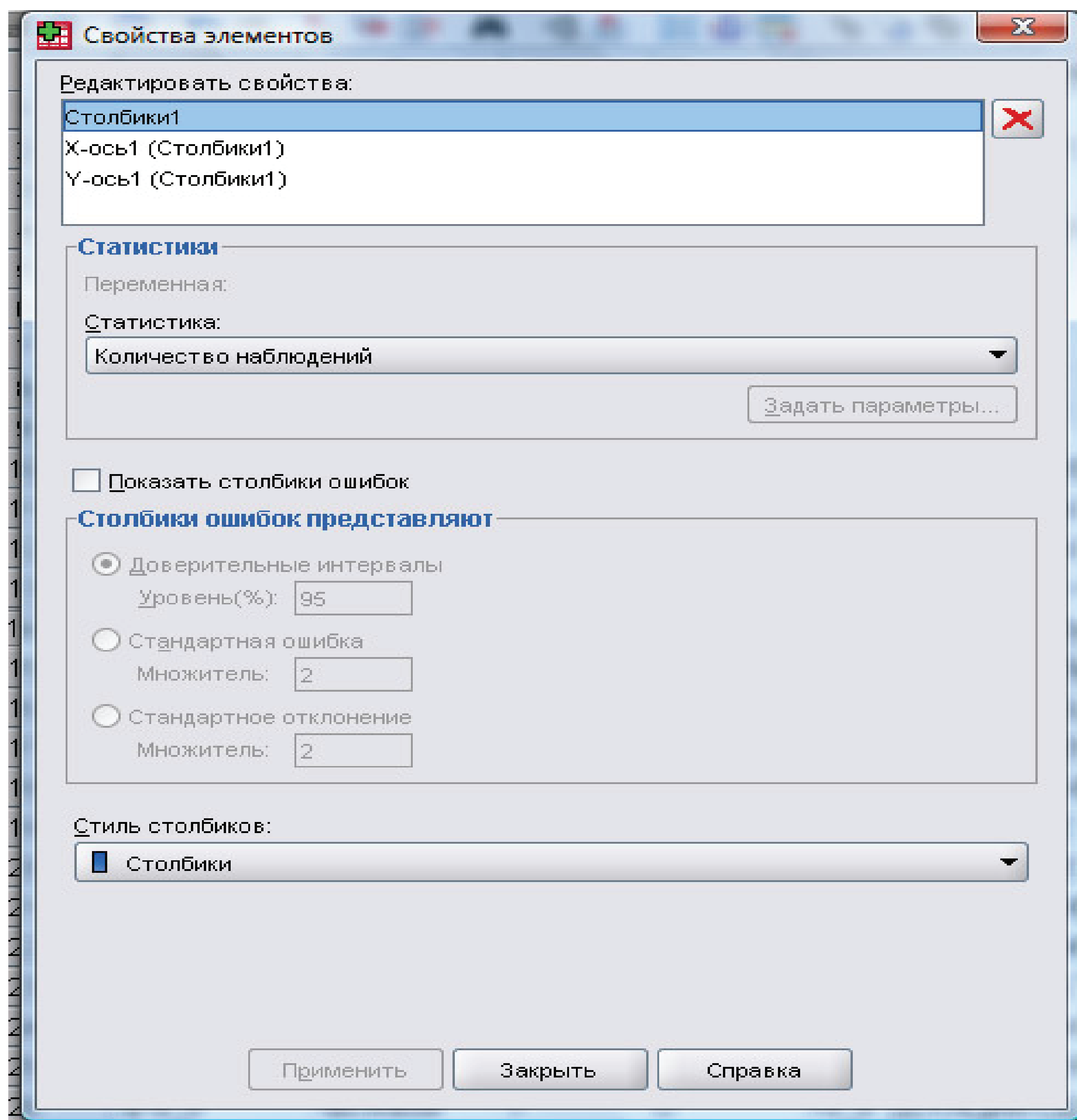


Рис. 4.6. Диалоговое окно «Свойства элементов» Конструктора диаграмм

Далее, чтобы настроить диаграмму, придать ей тот вид, который нам нужен, поработаем со свойствами элементов диаграммы в диалоговом окне «Свойства элементов» (рис. 4.6). В верхнем разделе окна выберем Ось X для того, чтобы настроить ее вид. Для начала исключим из графика тех респондентов, которые не дали ответ на этот вопрос. Выберем этот пункт в списке вариантов ответов и нажмем кнопку «удалить» (в виде креста красного цвета) справа от окна со списком вариантов (рис. 4.7). Отметим пункт «Не показывать пустые категории», нажмем кнопку «Применить» внизу окна, а затем кнопку «Закрыть».

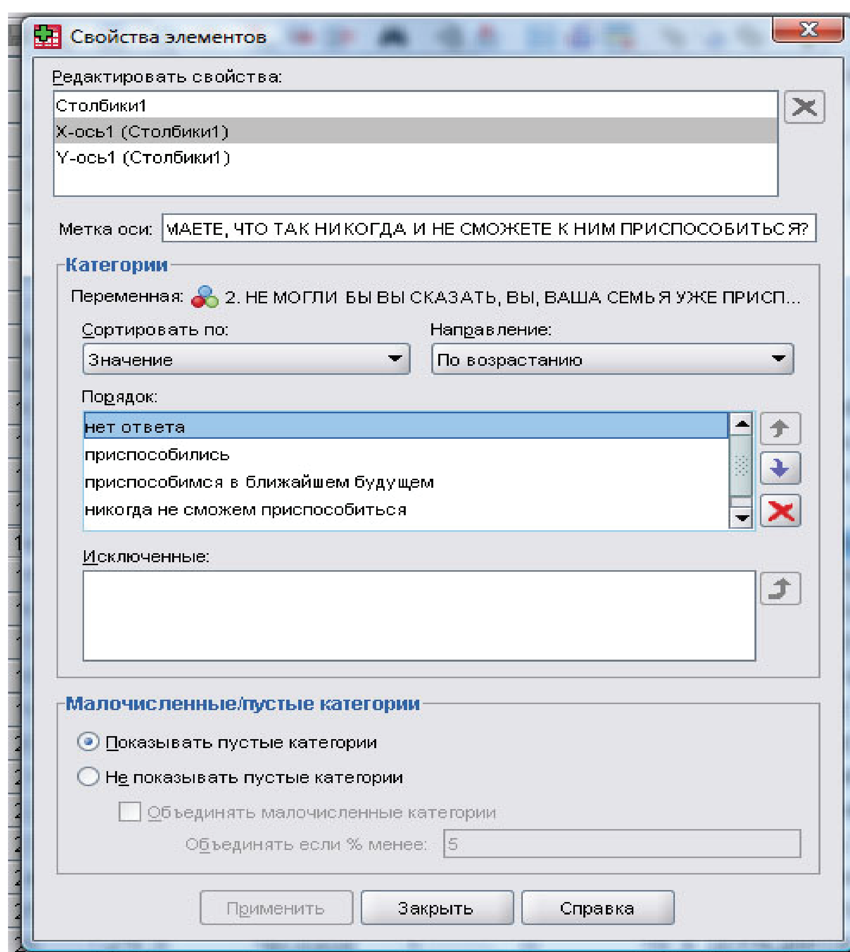


Рис. 4.7. Настройка переменной для оси X с помощью диалогового окна «Свойства элементов»

Далее в конструкторе диаграмм нажмем кнопку «ОК», чтобы получить диаграмму в окне вывода (рис. 4.8).

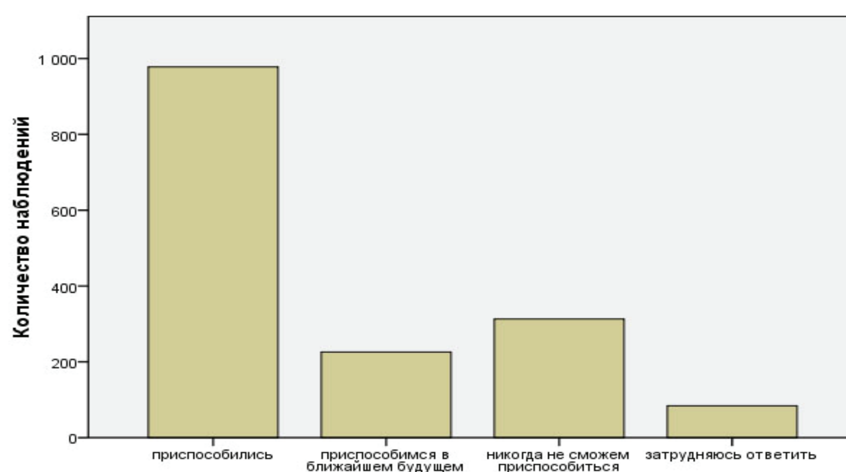


Рис. 4.8. Столбчатая диаграмма с результатами распределения ответов на вопрос о приспособленности россиян к переменам, произошедшим в стране за последние 10 лет

Показанная на рисунке диаграмма отражает результаты распределения на вопрос в частотном отношении, которое можно легко заменить на процентное, заменив в разделе «Статистика» окна «Свойства элементов» позицию «Количество наблюдений» на «Проценты».

Аналогичным образом для этого же примера составим круговую диаграмму (рис. 4.9). В данном случае показательны оба вида диаграмм, однако круговая диаграмма все-таки более наглядно представляет долю каждой группы респондентов в общей их массе. Какой вид диаграммы использовать в каждом конкретном случае, исследователь решает сам исходя из своих задач и предпочтений.

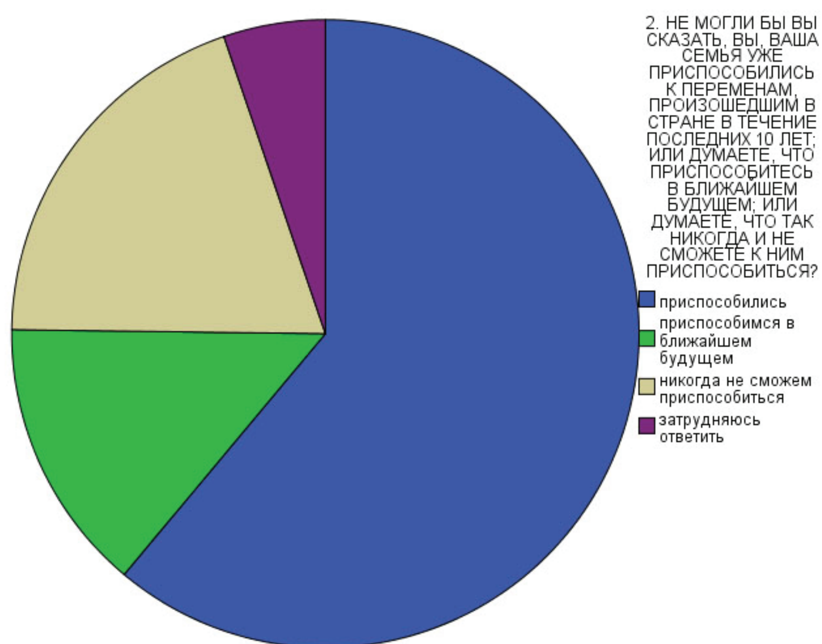


Рис. 4.9. Круговая диаграмма с результатами распределения ответов на вопрос о приспособленности россиян к переменам, произошедшим в стране за последние 10 лет

На этом этапе анализа интересные данные могут быть получены и благодаря анализу простейших статистических характеристик изучаемых переменных – средних величин (среднее арифметическое, медиана, мода), мер разброса (дисперсия, вариационный размах), квартилей и т.п. SPSS представляет исследователю широкие возможности для этого. Приведем примеры.

Для того, чтобы более наглядно представить возможности этих статистических показателей, воспользуемся переменной, измеренной по метрической шкале. В нашем случае это будет переменная, содержащая данные о среднедушевом доходе респондентов – qD9_2. Выберем ее в диалоговом окне «Частоты» (рис. 4.2). Далее воспользуемся функцией «Статистики».

Перед нами появится новое диалоговое окно (рис. 4.10). Для того чтобы сложилось представление о сути этих статистических показателей, отметим все пункты и нажмем «Продолжить», а затем «ОК». Единственное, что необходимо проделать перед этим – сделать отказы от ответов пропущенным значением, чтобы присвоенное этой категории ответов значение (закодированы как 999) не вносило путаницу в расчеты статистических показателей.

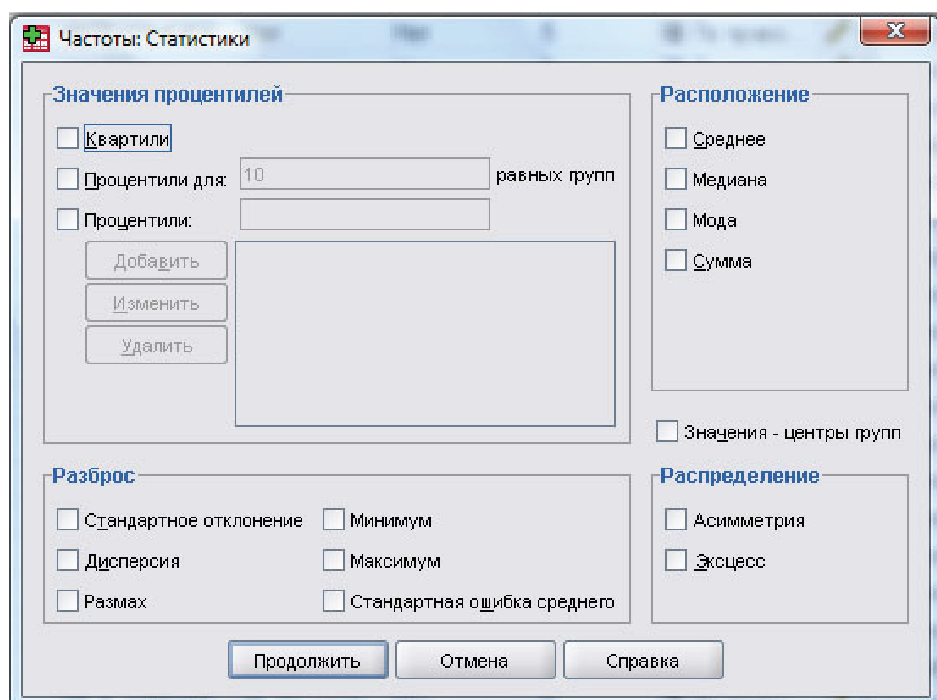


Рис. 4.10. Диалоговое окно «Частоты: Статистики»

Теперь мы получили таблицу 4.3 с разнообразными статистическими характеристиками интересующей нас переменной. Эта таблица напоминает таблицу 4.1, где мы уже видели число валидных и пропущенных значений изучаемой переменной. Однако в данном случае число валидных значений — 1332, а число пропущенных — 269. Кроме этого, таблица содержит дополнительно целый ряд показателей.

Первая группа этих показателей – различные средние. Среднее арифметическое, определенное как сумма значений, деленная на их количество, обычно позволяет исследователю судить о среднем уровне доходов или расходов респондентов (как в целом, так связанные с конкретными видами занятости), средних затратах времени на те или иные виды деятельности и т.п. Поэтому логично использование среднего арифметического только с метрическими шкалами, а не с номинальными или ранговыми. В нашем случае среднее арифметическое со стандартной ошибкой среднего говорят о том, что среднедушевой доход респондентов составляет 5589,89 руб. ± 189,16 руб.

Таблица 4.3

Статистики процедуры «Частоты»

N	Валидные	1332
	Пропущенные	269
Среднее		5559,89
Стд. ошибка среднего		189,162
Медиана		4238,89[a]
Мода		5000
Стд. Отклонение		6903,754
Дисперсия		4,766E7
Асимметрия		17,644
Стд. ошибка асимметрии		,067
Эксцесс		475,829
Стд. ошибка эксцесса		,134
Размах		199917
Минимум		83
Максимум		200000
Сумма		7405774
Процентили	25	2997,08[b]
	50	4238,89
	75	6672,93

[a] Вычислено по сгруппированным данным.

[b] Процентили вычисляются по сгруппированным данным.

Медиана и мода также являются средними величинами, однако медиану можно использовать не только с метрическими шкалами, но и с ранговыми, а моду еще и с номинальными шкалами. Медиана делит изучаемую совокупность пополам, то есть одинаковое количество значений в массиве меньше медианы, и такое же — больше. В случае, когда медиана лежит между двумя значениями ряда — она равна их средней арифметической. В нашем случае медиана — 4238,89 руб. Это значит, что половина респондентов имеет среднедушевой доход менее этой величины, а вторая половина — больше этой величины. Мода же является значением, которое наиболее часто встречается в изучаемой выборке. В нашем случае — это самый популярный ответ — 5000 руб.

Следующая группа показателей представлена мерами разброса — стандартное отклонение, дисперсия, вариационный размах, минимум и максимум. Эти показатели позволяют судить о степени однородности изучаемой совокупности. Стандартное отклонение представляет собой показатель, равный квадратному корню из другой меры разброса — дисперсии. Дисперсия, в свою очередь, равна сумме квадратов отклонений всех измеренных значений от их среднего значения, деленная на количество измерений (т. е. коли-

чество наблюдений). Большое значение дисперсии и стандартного отклонения свидетельствуют о большой неоднородности изучаемой совокупности. Однако зачастую объяснить напрямую, что именно скрывает за собой конкретное значение дисперсии, невозможно. В нашем случае можно констатировать, что стандартное отклонение в 6903 руб. при среднем значении 5559 руб. следует признать существенным. Это свидетельствует о значительной неоднородности изучаемой совокупности по среднедушевому доходу. На это же указывают и минимальное (83), и максимальное (200000) значения и, соответственно, разница между ними — вариационный размах (199917).

Показатели симметричности распределения (асимметрия и эксцесс) показывают насколько симметрично расположены ответы респондентов относительно среднего значения. Асимметрия учитывает равномерность распределения, то есть насколько одинаковое количество значений располагается по обе стороны выборки от среднего значения. Если коэффициент равен нулю, то это говорит о том, что перед нами абсолютно симметричное (нормальное) распределение. В нашем случае асимметрия отрицательная, так как значение асимметрии большое. А большое значение эксцесса указывает, что распределение является пологим.

Разделение исследуемой совокупности респондентов на квартили (или процентили) позволяет выделить те рубежи, за которыми лежит определенное число респондентов. Например, если мы пользуемся квартилями (как в нашем примере), то мы можем видеть, что 25% имеют среднедушевой доход менее 2997 руб., 50% — менее 4238 руб., а 75% — менее 6672 руб. Уже само по себе такое распределение весьма показательно. Если мы соотнесем его со значением среднего (5559 руб.), то увидим, что более половины респондентов имеют среднедушевой доход ниже среднего. Но сколько именно — из квартилей не ясно. Для того чтобы понять это, воспользуемся разделением совокупности на равные группы по процентилям — зададим значение «10%» и получим, что число людей, имеющих доход ниже среднего находится в пределах от 60 до 70% изучаемой совокупности (см. табл. 4.4).

Хотя из такого распределения тоже нельзя узнать точное число людей, среднедушевой доход которых не превышает среднеарифметического, но уже ясно, что они составляют около двух третей населения. А это уже говорит о многом. Также можно сравнить доходы 10% самых бедных (получают менее 1 750 руб.) и самых богатых (получают более 10 000 руб.), и получится, что доходы самых бедных и самых богатых разнятся в 5,7 раза. На основании этого становится ясна степень неоднородности доходов населения. В каждом конкретном случае исследователь решает, как именно делить исследуемую совокупность, если это вообще необходимо.

Таблица 4.4

Среднедушевой доход респондентов

N	Валидные	1332
	Пропущенные	269
Процентили	10	1750,00
	20	2500,00
	25	3000,00
	30	3000,00
	40	3500,00
	50	4250,00
	60	5000,00
	70	6000,00
	75	6667,00
	80	7500,00
	90	10000,00

Из приведенных выше примеров становится ясно, в общих чертах, как и для чего применяется частотный анализ в социологии. Однако прежде, чем перейти к следующему этапу анализа – двухмерному анализу данных – следует остановиться еще на одном важном вопросе.

До сих пор речь велась в основном о примерах альтернативных вопросов, на которые респондент не может дать сразу несколько ответов. Однако на практике социолог может столкнуться с необходимостью использования и других вопросов – неальтернативных, на которые респондент может дать несколько ответов одновременно. В этом случае построение частотных таблиц несколько отличается от описанного выше алгоритма.

В SPSS существует два способа кодирования и обработки неальтернативных вопросов – дихотомический и категориальный. В первом случае каждый вариант ответа кодируется как отдельная переменная с вариантами ответа 1 – «Признак присутствует» и 0 – «Признак отсутствует». Во втором случае в одной переменной кодируются значения всех вариантов ответов, а затем эта переменная дублируется необходимое количество раз. Совершенно очевидно, что в этом случае должно быть заранее известно максимальное количество возможных ответов, поэтому второй способ используется только тогда, когда в самом вопросе заложено ограничение типа «отметьте не более … вариантов».

Принципиальной разницы между этими двумя способами кодирования множественных ответов респондентов нет, хотя и существуют определенные различия. Поэтому в каждом случае исследователь сам решает, каким из возможных способов воспользоваться. Здесь же, не вдаваясь в общее и особенное этих двух методов, опишем специфику построения частотных таблиц для таких вопросов.

При дихотомическом методе у исследователя в распоряжении имеется несколько переменных (число их соответствует всем возможным вари-

антам ответов на неальтернативный вопрос). Для построения частотной таблицы необходимо объединить эти переменные в общий набор переменных. Для этого воспользуемся командой «Задать наборы переменных» в пункте «Множественные ответы» из меню «Анализ» (рис. 4.11).

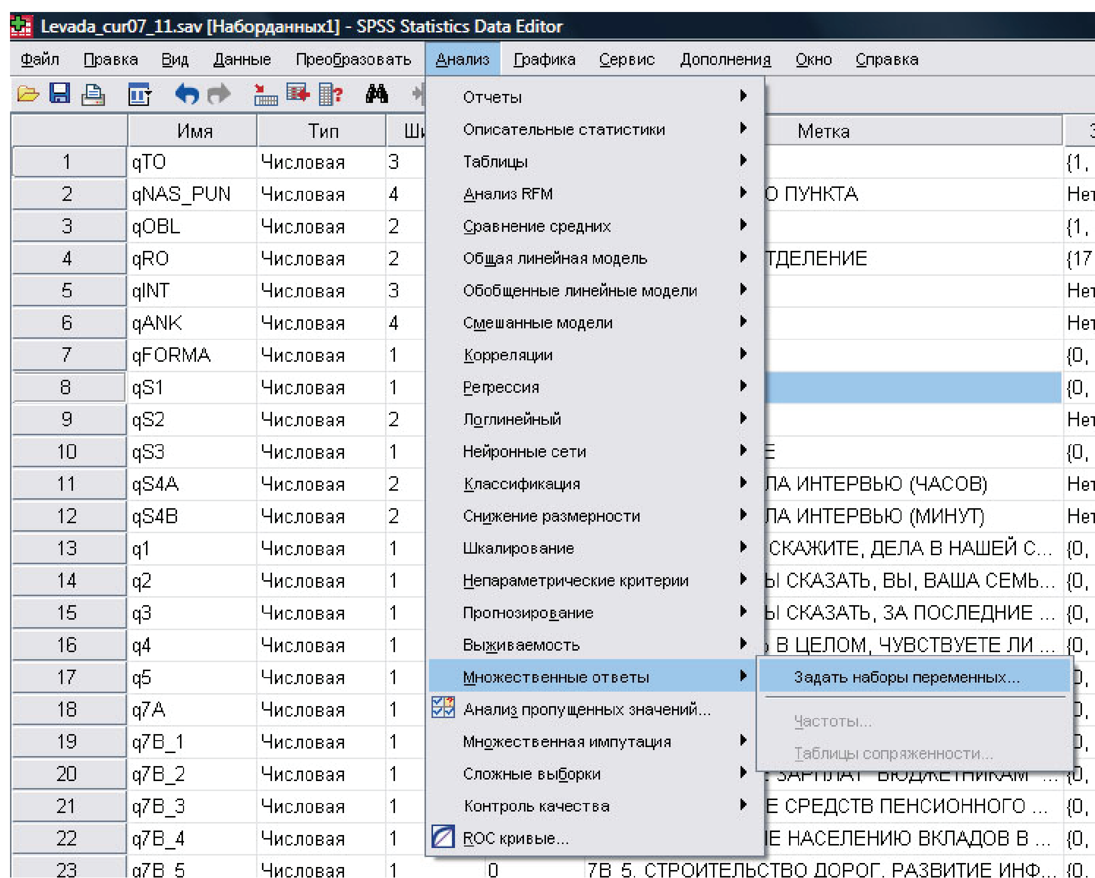


Рис. 4.11. Определение набора переменных с множественными ответами с помощью меню «Анализ: Множественные ответы: Задать набор переменных»

После выбора этой команды перед вами появится диалоговое окно, изображенное на рисунке 4.12.

В левом поле этого окна приводится список доступных переменных, в следующем за ним поле – переменные в наборе. Первоначально это поле является пустым, как и остальные. В нашем примере мы уже перенесли часть однотипных переменных для набора. Теперь следует выбрать тот из двух способов кодирования, который применен нами. В нашем случае мы оставляем дихотомический способ и указываем, что следует учитывать значение, закодированное как «1». Ниже определим имя будущей переменной и присвоим ей метку значения. После этого нажимаем кнопку «Добавить». Теперь в крайнем правом поле диалогового окна появился выделенный нами набор множественных ответов. Нажмем кнопку «Закрыть» и перейдем к построению частотной таблицы.

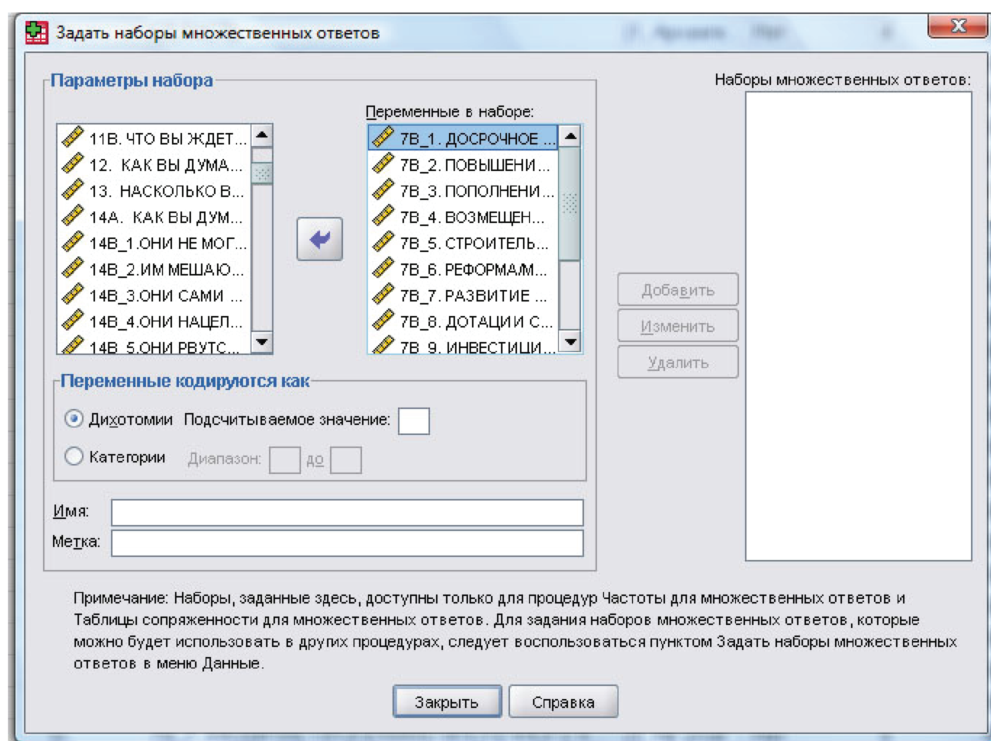


Рис. 4.12. Диалоговое окно «Задать наборы множественных ответов»

Для построения частотной таблицы перейдем в пункт «Множественные ответы» в меню «Анализ» и выберем ставшую доступной позицию «Частоты». Результатом будет появление диалогового окна «Частоты для множественных ответов» (рис. 4.13), которым мы воспользуемся для построения частотной таблицы.

Для вывода частотной таблицы перенесем имеющийся набор множественных ответов из левого поля в правое и нажмем «ОК». Таблица, которую мы получим, будет иметь следующий вид (см. таблицу 4.5). В данной таблице для всех наблюдаемых частот выводятся два различных процентных значения – процент от общего числа данных ответов и процент от общего числа всех наблюдений. Сумма процентов от общего числа ответов всегда будет равна 100%, тогда как сумма от общего числа наблюдений будет зависеть от числа данных респондентами ответов. Для нашего примера это 269,1%, что говорит о том, что каждый респондент в среднем отметил 2,7 возможных вариантов ответа.

При использовании категориального метода кодирования множественных ответов построение частотных таблиц принципиально не отличается. Для выполнения этой процедуры следует воспользоваться тем же сервисом (см. рис. 11–12). Однако в диалоговом окне, изображенном на рисунке 12, следует не только выбрать категориальный метод, но и отметить границы интервала, в которые будут вписаны множественные ответы. Например, если тот же вопрос о возможных вариантах использования излишков стабилизационного фонда закодировать категориально, то получится вот что:

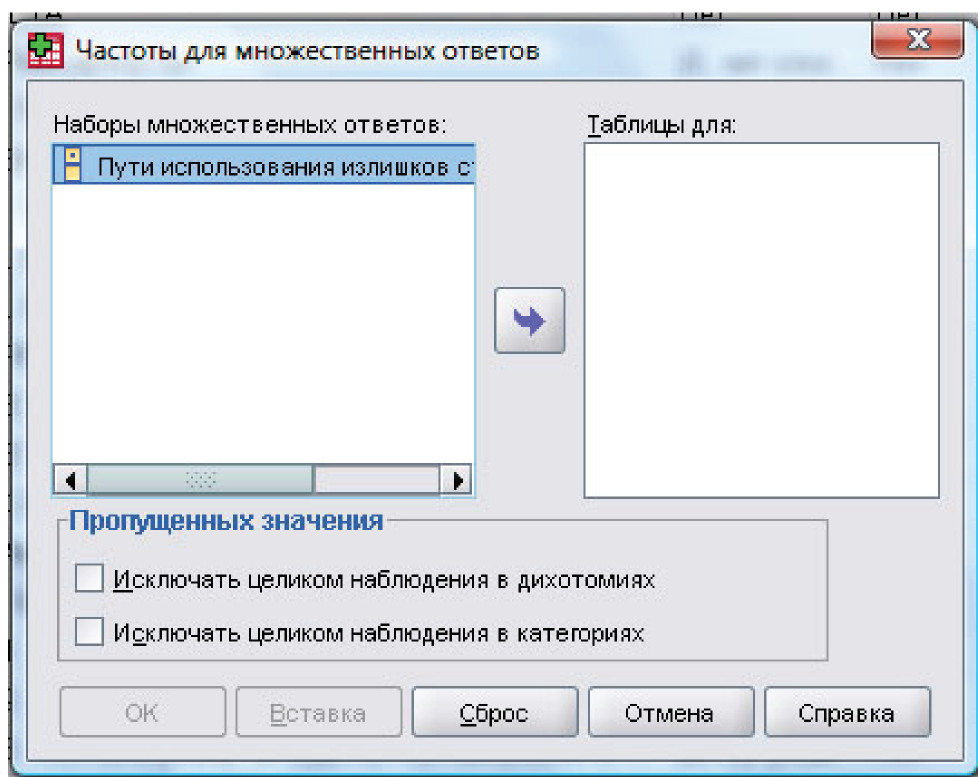


Рис. 4.13. Диалоговое окно «Частоты для множественных ответов»

1. Досрочное погашение внешнего долга;
2. Повышение зарплат «бюджетникам»;
3. Пополнение средств Пенсионного фонда;
4. Возмещение населению вкладов в Сбербанке, обесценившихся в 1992 г.;
5. Строительство дорог, развитие инфраструктуры;
6. Реформа/модернизация ЖКХ;
7. Развитие программы ипотечного кредитования;
8. Дотации сельскому хозяйству;
9. Инвестиции в перспективные экономические проекты/отрасли промышленности;
10. Модернизация армии;
11. Развитие образов-я, здравоохр., науки и культуры;
12. Снижение налогового бремени на бизнес;
13. Другое;
14. Затрудняюсь ответить.

В этом случае при определении набора множественных ответов можно задать границы интервала от 1 до 14. Однако это не обязательно. Если нам нужны только содержательные ответы, то следует задать границы интервала от 1 до 13, а если не нужна и категория «Другое», то можно исключить и ее, задав границы от 1 до 12 и т.д. Такая свобода действий в очередной раз позволяет убедиться в широких возможностях, которые предоставляет SPSS исследователю.

Таблица 4.5

$q7b_gr Частоты

		Ответы		Процент наблюдений
		N	Процент	
Пути использования излишков стаб. фонда[a]	7B_1. ДОСРОЧНОЕ ПОГАШЕНИЕ ВНЕШНЕГО ДОЛГА	139	3,2%	8,7%
	7B_2. ПОВЫШЕНИЕ ЗАРПЛАТ "БЮДЖЕТНИКАМ"	559	13,0%	34,9%
	7B_3. ПОПОЛНЕНИЕ СРЕДСТВ ПЕНСИОННОГО ФОНДА	596	13,8%	37,2%
	7B_4. ВОЗМЕЩЕНИЕ НАСЕЛЕНИЮ ВКЛАДОВ В СБЕРБАНКЕ, ОБЕСЦЕНИВШИХСЯ В 1992 ГОДУ	283	6,6%	17,7%
	7B_5. СТРОИТЕЛЬСТВО ДОРОГ, РАЗВИТИЕ ИНФРАСТРУКТУРЫ	395	9,2%	24,7%
	7B_6. РЕФОРМА/МОДЕРНИЗАЦИЯ ЖКХ	320	7,4%	20,0%
	7B_7. РАЗВИТИЕ ПРОГРАММЫ ИПОТЕЧНОГО КРЕДИТОВАНИЯ	219	5,1%	13,7%
	7B_8. ДОТАЦИИ СЕЛЬСКОМУ ХОЗЯЙСТВУ	479	11,1%	29,9%
	7B_9. ИНВЕСТИЦИИ В ПЕРСПЕКТИВНЫЕ ЭКОНОМИЧЕСКИЕ ПРОЕКТЫ/ ОТРАСЛИ ПРОМЫШЛЕННОСТИ	232	5,4%	14,5%
	7B_10.МОДЕРНИЗАЦИЯ АРМИИ	168	3,9%	10,5%
	7B_11.РАЗВИТИЕ ОБРАЗОВ-Я, ЗДРАВООХР., НАУКИ И КУЛЬТУРЫ	691	16,0%	43,2%
	7B_12.СНИЖЕНИЕ НАЛОГОВОГО БРЕМЕНИ НА БИЗНЕС	70	1,6%	4,4%
	7B_13.ДРУГОЕ	30	,7%	1,9%
	7B_14.ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	127	2,9%	7,9%
Всего		4308	100,0%	269,1%

[a] Дихотомическая группа подсчитывается по значению 1.

Вопросы и задания

1. Дайте общую характеристику частотного анализа данных и его возможностей.
2. Опишите формы графического изображения и принципы построения столбчатых диаграмм.
3. Какие статистические показатели можно получить при применении частотного анализа в SPSS?
4. Дайте определение понятия «меры разброса». Приведите примеры мер разброса и их интерпретации.
5. О чем говорит нулевое значение асимметрии данных при частотном анализе?
6. В чем отличие между дихотомическим и категориальным способом обработки вопросов с множественными ответами в SPSS?

Список литературы

1. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей/А. Бююль, П. Цёфель. — СПб.: ДиаСофтЮП, 2005. — 608 с.

2. Венецкий, И.Г. Вариационные ряды и их характеристики/И.Г. Венецкий. — М.: Статистика, 1970. — 159 с.

3. Крыштановский, А.О. Анализ социологических данных / А.О. Крыштановский — М.: Изд-во «ГУ ВШЭ», 2007. — 281 с.

4. Фишер, Р.А. Статистические методы для исследователей/Р.А. Фишер. — М.: Госстатиздат, 1958. — 326 с.

Лекция 5. Таблицы сопряженности

Многочисленный цифровой материал, имеющийся в распоряжении социолога, для удобного сопоставления и анализа, выявления определенных закономерностей следует оформить в виде таблиц. Использование таблиц в статистическом и социологическом анализе – насущная необходимость. Ведь без их помощи невозможно не запутаться в бесконечном числе цифр и показателей. Таблицы же помогают сгруппировать материал и выделить в нем наиболее важные и интересные показатели. По содержанию таблицы делятся на аналитические и неаналитические. В неаналитических таблицах помещаются необработанные статистические данные, необходимые лишь для информации и констатации. Как правило, это одномерные таблицы, рассмотренные выше. Аналитические таблицы являются результатом обработки, группировки и анализа цифровых показателей. Поэтому их еще называют таблицами сопряженности.

Важная роль в статистическом анализе принадлежит таблицам сопряженности. Именно эти таблицы позволяют из большого массива разрозненных данных выбрать и увидеть имеющиеся в нем взаимосвязи. Благодаря использованию этих таблиц исходная совокупность статистических данных разбивается на группы, каждая из которых объединена сходными характеристиками, выраженными в общих показателях. Кроме того, таблицы сопряженности не только позволяют констатировать наличие или отсутствие связи между изучаемыми переменными, но и компактно и вполне наглядно представить данные, что также немаловажно для социолога.

Если до появления современных средств анализа составление таких таблиц и их статистическое изучение требовало от исследователя социальных процессов немалых усилий, то теперь с помощью программы SPSS этот процесс существенно упростился.

С помощью процедуры «Таблицы сопряженности» в SPSS можно не только сформировать двумерные и многомерные таблицы, но и без трудоемких утомительных процедур вычислить целый ряд статистических критериев и мер силы связи для двумерных таблиц.

Для начала попробуем создать из имеющейся базы данных простую таблицу сопряженности двух признаков. Возьмем для примера переменные «Пол» и «Уверенность в завтрашнем дне» и попробуем с помощью таблицы сопряженности выяснить – что нам может дать такая группировка и есть ли взаимосвязь между этими показателями.

Чтобы построить таблицу сопряженности с помощью программы SPSS, необходимо в меню «Анализ» выбрать пункт «Описательные статистики», а в нем – «Таблицы сопряженности» (см. рис. 5.1).

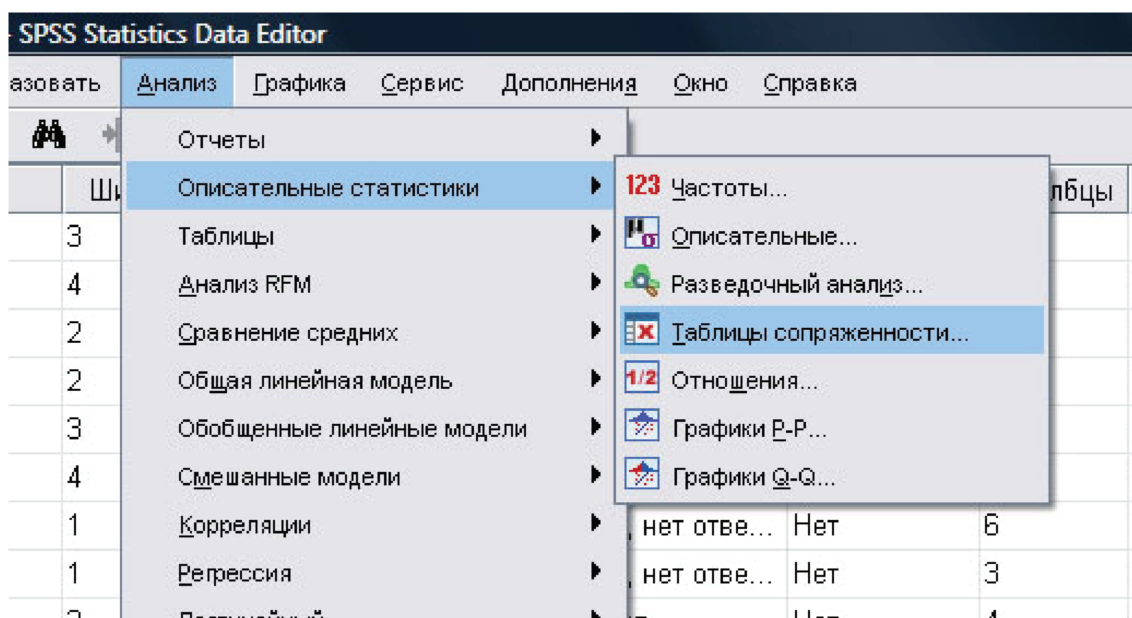


Рис. 5.1. Построение таблиц сопряженности с помощью меню «Анализ: Описательные статистики: Таблицы сопряженности»

Проделав эти действия, вы увидите на мониторе своего компьютера диалоговое окно следующего вида (см. рис. 5.2).

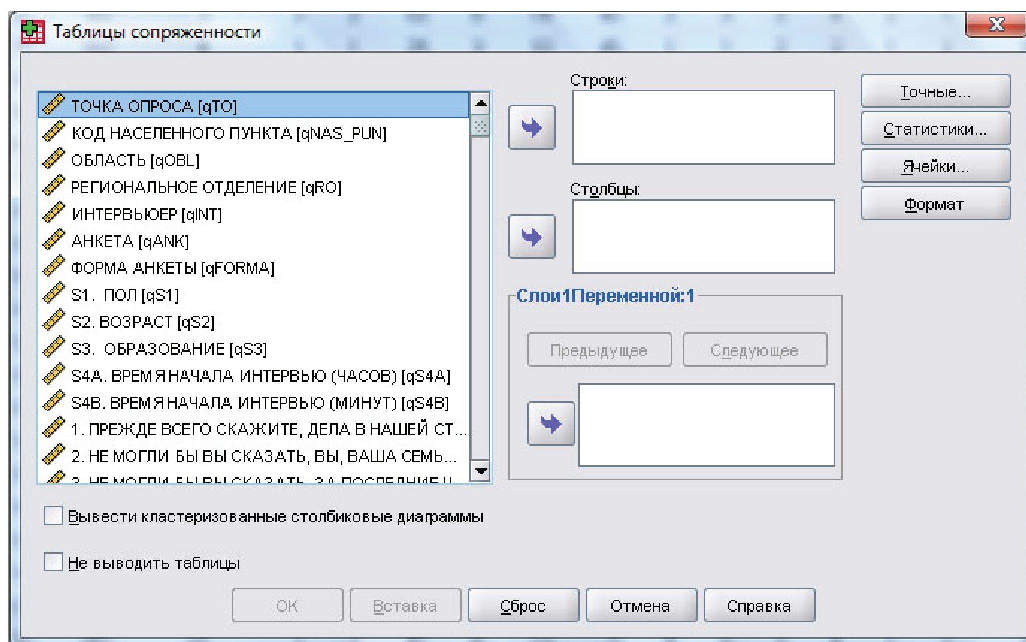


Рис. 5.2. Диалоговое окно «Таблицы сопряженности»

В левой части появившегося диалогового окна виден список доступных для анализа переменных, в правой его части имеются три пустых поля с соответствующими заголовками «Строки», «Столбцы» и «Слои», а также функциональные кнопки «Точные…», «Статистики…», «Ячейки…» и

«Формат…». Для того чтобы построить таблицу сопряженности для выбранных нами переменных, найдем в левой части окна и выделим в списке доступных переменных переменную «Пол». Затем с помощью щелчка левой кнопкой мыши по соответствующей стрелке перенесем ее в поле «Строки» (она будет строковой переменной), а переменную «Уверенность в завтрашнем дне» в поле «Столбцы» (она будет столбцовой переменной). Результаты этих действий видны на рисунке 5.3.

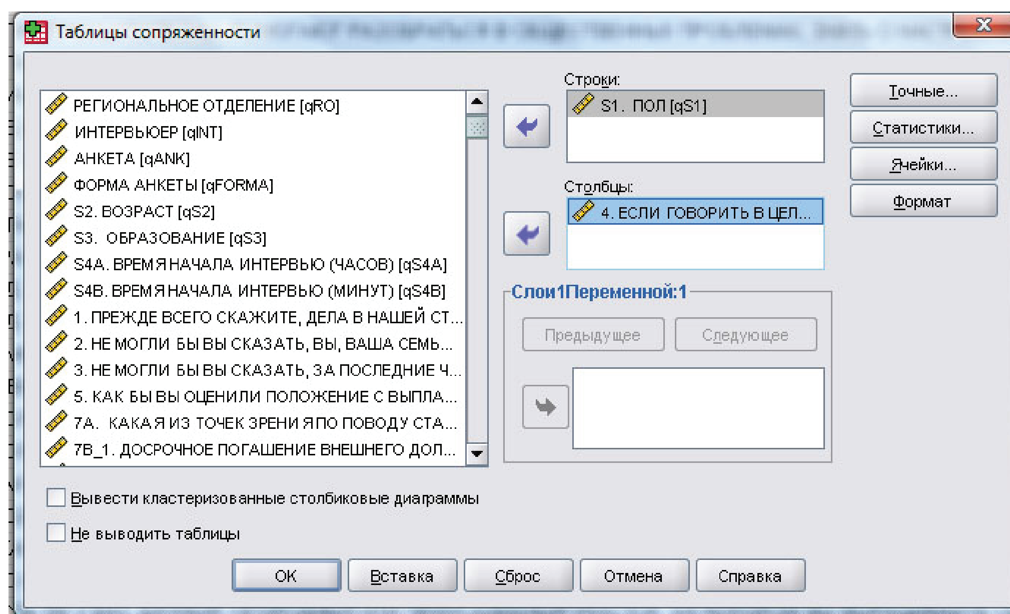


Рис. 5.3. Диалоговое окно «Таблицы сопряженности» с внесенными строковой и столбцовой переменными

Здесь сразу следует отметить один важный момент. Как вы могли убедиться, программа SPSS с помощью данного диалогового окна открывает широкие возможности по моделированию таблиц сопряженности. Говоря проще, программа позволяет легко менять одни переменные на другие, строковые переменные на столбцовые и наоборот. Однако исследователь должен четко представлять, что ему нужно от программы и, что еще более важно, знать – как правильно выбрать строковую и столбцовую переменные.

Помните: как правило, показатели, которые характеризуются в таблице (подлежащее таблицы), должны быть строковыми переменными, а логический предмет таблицы – то есть данные, которые характеризуют подлежащее (сказуемое таблицы) – столбцовыми переменными. Так и в нашем случае: подлежащее таблицы (то есть переменная, которую мы хотим изучить) – пол, а сказуемое таблицы (то есть данные, которые характеризуют подлежащее, в данном случае пол респондентов) – уверенность в завтрашнем дне. В такой таблице сопряженности мы можем увидеть зависимость уверенности респондентов в завтрашнем дне от их пола. Построение табли-

цы сопряженности со строковой переменной «Уверенность в завтрашнем дне» и столбцовой переменной «Пол» с логической точки зрения было бы ошибочным, хотя и возможным в программе SPSS.

После выбора строковой и столбцовой переменной нажмем «ОК» и получим таблицы следующего вида (см. таблицы 5.1 и 5.2).

Таблица 5.1

Сводка обработки наблюдений

	Наблюдения					
	Валидные		Пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ? * S3. ОБРАЗОВАНИЕ	1601	100,0%	0	,0%	1601	100,0%

Таблица 5.2

Таблица сопряженности переменных «ПОЛ» и «ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?»

		4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?					Итого
		определенно да	скорее да	скорее нет	определенно нет	затрудняюсь ответить	
S1. ПОЛ	мужской	78	262	227	133	25	725
	женский	59	229	332	205	51	876
Итого		137	491	559	338	76	1601

Первая представляет собой общую сводку обработки наблюдений, из которой видно, что в массиве 100%, или 1601 валидное значение и ни одного пропущенного. Вторая же – результат группировки по двум выбранным нами переменным, или собственно таблица сопряженности этих переменных. Как видим, значения переменной «Пол» выведены в строках, а переменной «Уверенность в завтрашнем дне» – в столбцах. Такая группировочная таблица уже дает некоторые возможности для анализа данных.

Как видим из таблицы 2, если 78 мужчин выразили определенную уверенность в завтрашнем дне, то лишь 59 женщин сделали то же самое. При этом на очевидную неуверенность в будущем указали только 133 мужчины из 725 и 205 женщин из 876 опрошенных. К сожалению, из этих абсолютных частот, выведенных в таблицу по умолчанию, нельзя достоверно судить о масштабах взаимосвязи переменных «Пол» и «Уверенность в завтрашнем дне» и тем более о масштабах разницы в отношении собственной уверенности у мужчин и женщин, хотя уже приведенная таблица зафиксировала в

общем виде некоторые отличия в ответах респондентов. При дальнейшем анализе они могут либо подтвердиться, либо будут опровергнуты и признаны статистически незначимыми (случайными).

Еще одним подтверждением нашего предположения о наличии взаимосвязи между полом респондента и степенью его уверенности в завтрашнем дне служит таблица 5.3, в которой, наряду с частотами, наблюдаемыми в совокупности опрошенных (выявленных в ходе опроса), приводятся частоты теоретически ожидаемые – при условии соответствия нормальному распределению. Они равны произведению сумм соответствующих строки и столбца, деленному на общую сумму частот в массиве данных. Например, ожидаемое число мужчин, выражающих определенную уверенность в завтрашнем дне, равно 62 = (725 × 137) / 1601, а число женщин, определенно неуверенных, равно 184,9 = (876 × 338) / 1601. Причем ожидаемые частоты вовсе не обязательно могут быть целыми числами, ведь это теоретически рассчитанная вероятность совпадения. Хотя их можно округлить до целого.

SPSS все это вычисляет автоматически. Для этого достаточно в уже знакомом нам диалоговом окне (рис. 5.2) щелкнуть мышью на кнопке «Ячейки…». Перед вами появится диалоговое окно «Таблицы сопряженности: Вывод в ячейках» (см. рис. 5.4). В этом окне достаточно поставить галочку напротив не только наблюденных частот, но и ожидаемых, и нажать «ОК».

Таблица 5.3

Таблица сопряженности переменных «ПОЛ» и «ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?»

			4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?					Итого
			определенно да	скорее да	скорее нет	определенно нет	затрудняюсь ответить	
S1. ПОЛ	мужской	Частота	78	262	227	133	25	725
		Ожидаемая частота	62,0	222,3	253,1	153,1	34,4	725,0
	женский	Частота	59	229	332	205	51	876
		Ожидаемая частота	75,0	268,7	305,9	184,9	41,6	876,0
Итого		Частота	137	491	559	338	76	1601
		Ожидаемая частота	137,0	491,0	559,0	338,0	76,0	1601,0

Сравнение теоретически ожидаемых и фактически наблюдаемых частот в таблице 5.3 показывает, что теоретически ожидаемое число опрошенных мужчин, определенно уверенных в завтрашнем дне, превышает число

фактически наблюдаемых на 16, а число женщин – наоборот, меньше ожидаемого (теоретически рассчитанного) на это же значение. Исходя из этого можно заключить, что мужчины значительно увереннее чувствуют себя, и перспективы завтрашнего дня видятся им более радужными, чем женщинам.

Однако приведенные здесь расчеты не в полной мере позволяют представить масштабы взаимосвязи изучаемых переменных в рамках всего ансамбля опрошенных. Для этого нам понадобится перевести наши данные из абсолютных в относительные частоты (в проценты). Вообще, сразу следует оговориться, что в большинстве случаев различные сравнения данных для получения основных выводов будут делаться на основании именно относительных (процентных) показателей.

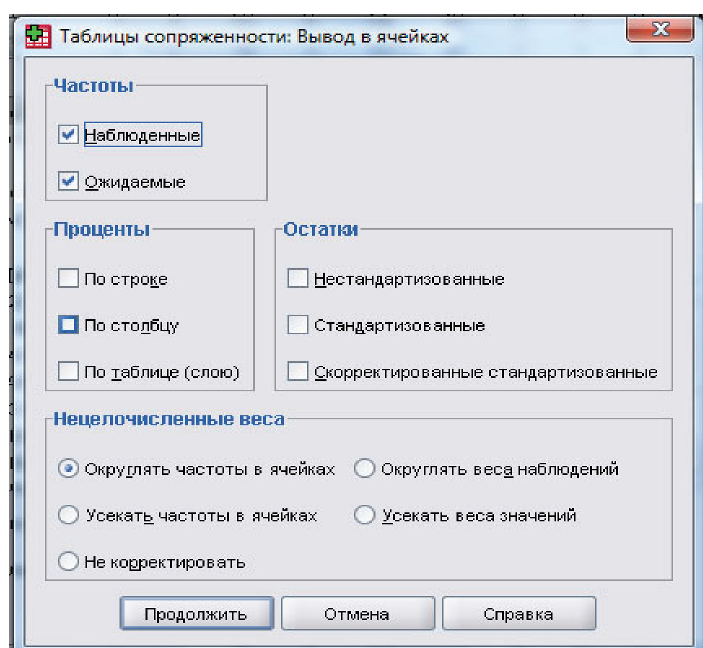


Рис. 5.4. Диалоговое окно «Таблицы сопряженности: Вывод в ячейках»

Чтобы увидеть в таблице не только абсолютные, но и относительные (процентные) значения, необходимо в уже знакомом нам диалоговом окне «Таблицы сопряженности: Вывод в ячейках» (рис. 5.4) проставить галочки напротив соответствующих окошек в разделе «Проценты». Здесь мы можем задать вывод процентов по строке, по столбцу и в целом по таблице. Если мы выберем первый вариант (проценты по строке), тогда за 100% будет взята сумма по строке, если по столбцу – то соответственно по столбцу. В том случае, если мы выберем позицию «по таблице» сумма всех валидных значений будет взята за 100% (в нашем случае это 1601 наблюдение). Можно выбрать и несколько позиций одновременно. В данном случае мы отметили все возможные позиции (по строке, по столбцу, по таблице) и получили таблицу следующего вида.

Теперь мы видим, что 78 мужчин, определенно уверенных в завтрашнем дне, это только 10% всех мужчин и лишь 5% из всех опрошенных. Но вместе с тем эти же 78 человек составляют большую часть (57%) всех определенно уверенных в завтрашнем дне. То же и с женщинами: 205 определенно неуверенных в завтрашнем дне женщин составляют почти четверть (23,4%) от их числа и, хотя это только 12,8% от всех опрошенных, зато более двух третей (67%) от всех определенно неуверенных. Как видим из данных таблицы 5.4, если определенная уверенность мужчин в завтрашнем дне по сравнению с женщинами не столь очевидна, то неуверенность женщин в сравнении с мужчинами – более чем.

Таблица 5.4

Таблица сопряженности переменных «ПОЛ» и «ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?»

			4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?					Итого
			определенно да	скорее да	скорее нет	определенно нет	затрудняюсь ответить	
S1. ПОЛ	мужской	Частота	78	262	227	133	25	725
		% в S1. ПОЛ	10,8%	36,1%	31,3%	18,3%	3,4%	100,0%
		% в 4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?	56,9%	53,4%	40,6%	39,3%	32,9%	45,3%
		% по таблице (слою)	4,9%	16,4%	14,2%	8,3%	1,6%	45,3%
	женский	Частота	59	229	332	205	51	876
		% в S1. ПОЛ	6,7%	26,1%	37,9%	23,4%	5,8%	100,0%
		% в 4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?	43,1%	46,6%	59,4%	60,7%	67,1%	54,7%
		% по таблице (слою)	3,7%	14,3%	20,7%	12,8%	3,2%	54,7%
Итого		Частота	137	491	559	338	76	1601
		% в S1. ПОЛ	8,6%	30,7%	34,9%	21,1%	4,7%	100,0%
		% в 4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% по таблице (слою)	8,6%	30,7%	34,9%	21,1%	4,7%	100,0%

Таким образом, изучение процентных показателей в перекрестной таблице позволяет сделать очень интересные выводы о взаимосвязи переменных, еще очевиднее указывая на большую уверенность мужчин в завтрашнем дне. В каждом конкретном случае, исходя из целей исследования, можно выбрать – какое именно вычисление процентов необходимо.

Приведенные выше примеры иллюстрируют самый первый – простейший уровень анализа таблиц сопряженности. Однако широчайшие возможности статистического пакета SPSS на этом не заканчиваются. Далее следует рассмотреть основные статистические возможности определения наличия и тесноты взаимосвязи между изучаемыми строковыми и столбцовыми переменными. В математической статистике существует множество способов для решения этой задачи. Мы же остановимся на наиболее важных и часто употребляемых из них.

Статистическая проверка наличия взаимосвязи. Хи-квадрат

Изучение таблиц сопряженности первым делом предполагает выяснение принципиальной возможности существования связей между рассматриваемыми переменными, и лишь затем, на основе применения формальных статистических критериев, выяснение силы этих связей.

Первое, что необходимо сделать для проверки гипотезы о наличии взаимосвязи рассматриваемых переменных – провести статистический тест хи-квадрат (χ^2). При проведении теста хи-квадрат проверяется нулевая гипотеза, согласно которой строковая и столбцовая переменные в таблице независимы, при этом косвенно проверяется зависимость обоих переменных (то есть в выявлении взаимосвязи между переменными мы идем от обратного). Две переменные считаются независимыми, если наблюдаемые частоты совпадают с ожидаемыми частотами.

В рассмотренном выше (на основе таблицы 5.3) примере наблюдаемые частоты отличаются от теоретически ожидаемых. Колебания первых и вторых для определенно уверенных в завтрашнем дне респондентов составляет 16, для скорее уверенных – 40 в пользу мужчин, для скорее неуверенных 26 и определенно неуверенных – 20 в пользу женщин. Как видим, наблюдаемые частоты отличаются от теоретически ожидаемых в каждой ячейке таблицы 5.3.

Казалось бы, такие существенные отличия между наблюдаемыми и ожидаемыми частотами – еще один аргумент в пользу наличия связи между переменными «Пол» и «Уверенность в завтрашнем дне». Однако чтобы получить этому математическое подтверждение, необходимо вычислить статистику коэффициента хи-квадрат, ведь без этого мы не можем быть уверены, что различия между наблюдаемыми и теоретически ожидаемыми частотами не являются случайными. Насколько случайны или закономерны эти разли-

чия на вскидку сказать нельзя, поэтому необходимо воспользоваться математическим инструментарием пакета программ SPSS.

Для того чтобы вычислить значение коэффициента хи-квадрат (χ^2) в уже знакомой нам последовательности команд «Анализ» – «Описательные статистики» – «Таблицы сопряженности» (рис. 5.1) выберем переменные «Пол» и «Уверенность в завтрашнем дне», как это делали ранее (рис. 5.3). Теперь в окне «Таблицы сопряженности: Вывод в ячейках» отметим не только наблюдаемые частоты, предлагаемые по умолчанию, но и ожидаемые. А в разделе «Остатки» отметим «Стандартизированные» и нажмем кнопку «Продолжить» (рис. 5.5).

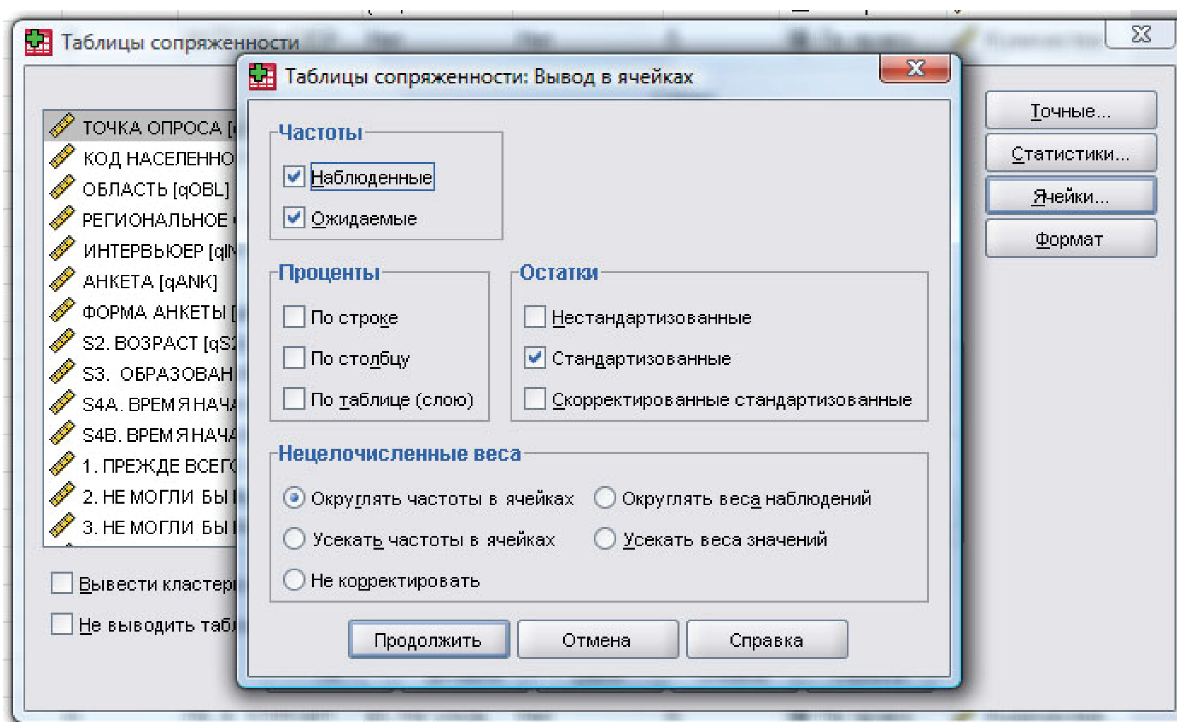


Рис. 5.5. Выбор вывода наблюдаемых и ожидаемых частот со стандартизированными остатками в диалоговом окне «Таблицы сопряженности: Вывод в ячейках»

Стандартизация остатков (разности наблюдаемых и ожидаемых частот) в данном случае необходима для того, чтобы привести к общему показателю переменные, имеющие различный вариационный размах и отличие значений на порядок. Только так мы сможем их сравнить и сделать правильные (статистически обоснованные) выводы. Стандартизированные значения остатков вычисляются программой SPSS как разность между наблюдаемыми и ожидаемыми значениями, деленная на квадратный корень из среднеквадратичного значения остатков. Значение стандартизированных остатков $\geq 2{,}0$ свидетельствует о достаточно значимом, $\geq 2{,}6$ – очень значи-

мом, ≥ 3,3 – сверхзначимом отклонении между наблюдаемыми и ожидаемыми частотами в рассматриваемой совокупности данных.

Далее, воспользовавшись кнопкой «Статистики…», увидим на экране диалоговое окно «Таблицы сопряженности: Статистики» (см. рис. 5.6), где выберем значение «Хи-квадрат» и нажмем кнопку «Продолжить», а затем «ОК».

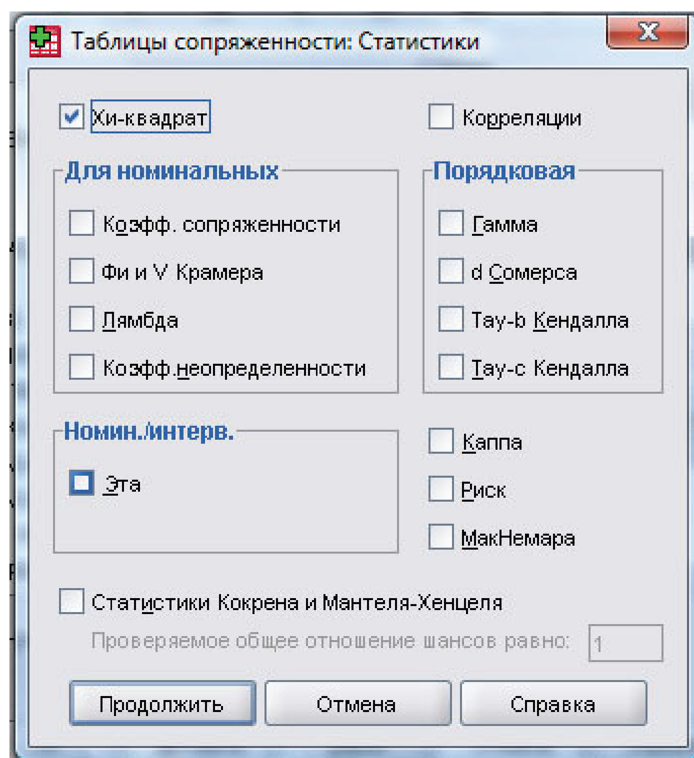


Рис. 5.6. Диалоговое окно «Таблицы сопряженности: Статистики»

После завершения этой процедуры в окне вывода программы SPSS получим две таблицы следующего вида (см. таблицы 5.5 и 5.6).

В первой из них дано соотношение наблюдаемых и ожидаемых частот со значениями стандартизированных остатков, а во второй – значение вычислений коэффициента хи-квадрат, по которому, в конечном счете, будет видно – есть статистически значимая взаимосвязь между изучаемыми переменными или нет.

Сразу необходимо отметить, что применение критерия хи-квадрат правомерно лишь при соблюдении двух очень важных условий. Первое из них заключается в том, что значение ожидаемых частот менее пяти не должно присутствовать более чем в 20% ячеек таблицы сопряженности (в нашем случае – не более чем в двух ячейках из десяти имеющихся). Второе условие: суммы по строкам и столбцам обязательно должны быть больше нуля во всех случаях. В нашем примере оба эти условия полностью соблюдены, о чем говорится в примечании к таблице 5.6, выведенном программой SPSS

автоматически. Учитывая соблюдение необходимых условий, применение коэффициента хи-квадрат вполне правомерно.

Таблица 5.5

Таблица сопряженности переменных «ПОЛ» и «ЕСЛИ ГОВОРИТЬВ ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?»

			4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?					Итого
			определенно да	скорее да	скорее нет	определенно нет	затрудняюсь ответить	
S1. ПОЛ	мужской	Частота	78	262	227	133	25	725
		Ожидаемая частота	62,0	222,3	253,1	153,1	34,4	725,0
		Стандартиз. остаток	2,0	2,7	-1,6	-1,6	-1,6	
	женский	Частота	59	229	332	205	51	876
		Ожидаемая частота	75,0	268,7	305,9	184,9	41,6	876,0
		Стандартиз. остаток	-1,8	-2,4	1,5	1,5	1,5	
Итого		Частота	137	491	559	338	76	1601
		Ожидаемая частота	137,0	491,0	559,0	338,0	76,0	1601,0

Таблица 5.6

Критерии хи-квадрат

	Значение	ст.св.	Асимпт. значимость (2-стор.)
Хи-квадрат Пирсона	34,876[a]	4	,000
Отношение правдоподобия	34,973	4	,000
Линейно-линейная связь	21,521	1	,000
Кол-во валидных наблюдений	1601		

[a] В 0 (,0%) ячейках ожидаемая частота меньше 5. Минимальная ожидаемая частота равна 34,42.

Как известно, если полученное значение статистики хи-квадрат велико, нулевая гипотеза отвергается – это означает, что между переменными имеется статистически значимая связь. Для того чтобы определить, достаточно ли велико наблюдаемое значение, его сравнивают с критической точкой Z, вычисленной по теоретическому распределению хи-квадрата. Если полученное в результате вычислений значение коэффициента хи-квадрат превышает критическое значение, то это свидетельствует в пользу несостоятельности нулевой гипотезы, об отсутствии взаимосвязи между рассматриваемыми переменными. Значения в критических точках зависят от числа степеней свободы, которые определяются числом строк и столбцов в таблице. Для таблицы RxC число степеней свободы равно числу строк (R) минус единица, умноженному на число столбцов (C) минус единица. В нашем случае: (2 – 1) × (5 – 1) = 4.

Значение коэффициента хи-квадрат для таблицы сопряженности переменных «Пол» и «Уверенность в завтрашнем дне» составило 34,876. Для таблицы с числом степеней свободы 4 при уровне значимости 0,01 в соответствии с таблицей критических точек распределения хи-квадрат это значение не должно превышать 13,3[11]. Как видим, вычисленное значение коэффициента хи-квадрат в нашем случае существенно превышает критическое значение из таблицы, поэтому нулевая гипотеза должна быть отвергнута.

Если при вычислениях коэффициента хи-квадрат вручную есть необходимость сравнивать его значение со значениями критических точек по специальным статистическим таблицам, то при вычислении коэффициента хи-квадрат с помощью статистического пакета SPSS этого не требуется. Здесь это было проделано лишь для примера.

Отсутствие необходимости использования таблицы критических точек коэффициента хи-квадрат существенно облегчает процедуру применения критерия хи-квадрат в социологическом исследовании. Значимость проведенных компьютером вычислений представлена в третьем столбце выведенной таблицы 5.6 – «Асимптоматическая значимость» (двухсторонняя).

В данном примере значение коэффициента хи-квадрат 34,876 для числа степеней свободы (ст.св.), равного 4, соответствует вероятности *p* для теоретического распределения меньше чем 0,000. Эта величина называется *наблюденный уровень значимости*. Если она достаточно мала (менее 0,05 или 0,01), нулевая гипотеза отвергается. У нас она ничтожно мала, поэтому нулевая гипотеза об отсутствии взаимосвязи между переменными «Пол» и «Уверенность в завтрашнем дне» отвергается. Как указывалось нами ранее, косвенным образом это является свидетельством фиксации статистически достоверного и значимого (а не случайного) наличия связи между указанными переменными.

Значения, выведенные в таблице 5.6 во второй и третьей строках, представляют собой расчеты значения коэффициента хи-квадрата Пирсона с поправкой на правдоподобие (строка 2) и поправкой Мантеля–Хензеля для линейно-линейных связей (строка 3). Это аналогичные обычному хи-квадрату коэффициенты, которые являются разновидностями статистических попыток стандартизации вычислений. При больших объемах выборочной совокупности, как и в нашем случае, они не отличаются существенно. Об особенностях их применения можно дополнительно прочитать в специальной литературе. Здесь лишь отметим, что коэффициент с поправкой Мантеля–Хензеля нельзя применять, если одна из шкал изучаемых переменных является шкалой наименований, или номинальной шкалой.

---

[11] См. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике / В.Е. Гмурман. – Изд. 3-е, перераб. и доп. – М., 1979. – С. 392.

Также в SPSS с помощью процедуры «Таблицы сопряженности» можно вычислить ряд показателей, основанных на коэффициенте хи-квадрат. Это коэффициенты сопряженности Пирсона *(C)*, Крамера, φ (фи), λ (лямбда) Гутмана, τ (тау) Гудмена–Краскэла и коэффициент неопределенности. Каждый из них имеет свои особенности. Выбрав все эти коэффициенты в диалоговом окне «Таблицы сопряженности: Статистики» (рис. 5.6) получим, кроме уже знакомых нам таблиц 5.5 и 5.6, таблицы 5.7 и 5.8. Здесь мы не будем подробно рассматривать их, отметим лишь самые общие моменты по особенностям применения и интерпретации, что немаловажно для социолога-исследователя.

Таблица 5.7

Направленные меры

			Значение	Асимптом. стданд. ошибка[a]	Прибл. T[b]	Прибл. значимость
Номинальная по номинальной	Лямбда	Симметричная	,049	,022	2,151	,031
		Зависимая S1. ПОЛ	,072	,033	2,078	,038
		Зависимая 4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?	,034	,021	1,584	,113
	Тау Гудмена и Краскала	Зависимая S1. ПОЛ	,022	,007		,000[c]
		Зависимая 4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?	,006	,002		,000[c]
	Коэффициент неопределенности	Симметричная	,010	,003	2,973	,000[d]
		Зависимая S1. ПОЛ	,016	,005	2,973	,000[d]
		Зависимая 4. ЕСЛИ ГОВОРИТЬ В ЦЕЛОМ, ЧУВСТВУЕТЕ ЛИ ВЫ УВЕРЕННОСТЬ В ЗАВТРАШНЕМ ДНЕ?	,008	,003	2,973	,000[d]

[a] Не подразумевая истинность нулевой гипотезы.

[b] Используется асимптотическая стандартная ошибка в предположении истинности нулевой гипотезы.

[c] На основании аппроксимации хи-квадрат.

[d] Вероятность отношения правдоподобия хи-квадрат.

Меры направленности связи – λ (лямбда) Гутмана, τ (тау) Гудмена–Краскэла и коэффициент неопределенности – позволяют зафиксировать возможность предсказать поведение одной (зависимой) переменной на основе поведения другой (независимой) переменной. Лямбда показывает, насколько велика ошибка предсказания. Чем ближе значение лямбды к нулю, тем хуже поведение одной переменной предсказывает поведение другой, и наоборот, чем ближе оно к единице, тем больше вероятность правильного прогноза.

Тем самым косвенно может быть выявлена взаимосвязь между переменными. Как видим из таблицы 5.7, в нашем примере значение лямбды слишком мало для того, чтобы оценивать поведение переменных и выявить, какая из них является зависимой. Поэтому программой автоматически было рассчитано среднее (симметричное) значение лямбды.

Симметричные меры связи, представленные в таблице 5.8, вычисляются на основе значения хи-квадрата с различными способами его стандартизации.

Коэффициент сопряженности Пирсона *C* представляет собой стандартизированный коэффициент хи-квадрат, значение которого, в отличие от самого хи-квадрата, меняется в пределах от 0 до 1, поэтому по его величине легче констатировать наличие или отсутствие взаимосвязи между переменными.

Таблица 5.8

Симметричные меры

		Значение	Прибл. значимость
Номинальная по номинальной	Фи	,148	,000
	V Крамера	,148	,000
	Коэффициент сопряженности	,146	,000
Кол-во валидных наблюдений		1601	

Существенным ограничением этого коэффициента является то, что из-за влияния на его величину суммы частот в таблице сопряженности его значение несопоставимо для разных (как правило, разноразмерных) таблиц, то есть с его помощью нельзя сравнить взаимосвязь двух признаков в одной таблице со связью других двух признаков в другой таблице.

Коэффициент сопряженности Крамера *V* так же принимает значение от 0 до 1, причем он может быть равен единице только в случае полного совпадения детерминантности признаков во всех наблюдаемых случаях. При этом, в случае если минимальное число строк или столбцов таблицы сопряженности, для которой он вычисляется равно двум (как в нашем случае), то значение этого коэффициента совпадет со значением коэффициента φ (фи) Фишера.

К сожалению, ни один из рассмотренных коэффициентов сопряженности не указывает напрямую на силу связи между рассматриваемыми переменными. Все они так или иначе лишь констатируют ее наличие или указывают на ее отсутствие, то есть если в одном случае, например, коэффициент Крамера составляет 0,2, а в другом – 0,3, то это вовсе не свидетельствует о том, что во втором случае связь между переменными сильнее. Это лишь является математическим свидетельством того, что во втором случае коэф-

фициент хи-квадрат с большей вероятностью значим, чем в первом, то есть больше вероятности того, что нулевая гипотеза об отсутствии взаимосвязи будет отвергнута.

В практике социологических исследований выявление наличия или отсутствия взаимосвязи между теми или иными рассматриваемыми переменными, как правило, недостаточно. Социологу не менее важно знать: насколько тесно взаимосвязаны изучаемые переменные, то есть насколько изменится одна, если изменится другая переменная. Вычисление значения коэффициента хи-квадрат для этого явно недостаточно.

Вычисление тесноты взаимосвязи. Коэффициенты корреляции

Для более детального изучения взаимозависимостей между переменными в социологии и ряде других наук применяются различные математико-статистические методы. Одним из широко применяемых методов является *метод корреляционного анализа*. Существующие коэффициенты корреляции позволяют определить и выразить количественно степень тесноты взаимосвязи одного признака с другим (явление парной корреляции) или с группой признаков (явление множественной корреляции).

В основе метода корреляционного анализа лежит идея сопоставления колебаний значений изучаемых признаков в зависимости друг от друга. Это можно проделать и в таблице сопряженности. Если в таблице колебание значений меняются по диагонали в том или ином направлении, то это чаще всего свидетельствует о наличии корреляции между признаками. Если же никакой закономерности в колебаниях признаков в таблице не прослеживается, то и корреляция неочевидна.

В таблицах 5.9 и 5.10 представлены результаты группировки респондентов по уровню доходов семьи в зависимости от уровня образования. При этом в таблице 5.9 графически выделена диагональ, к которой должны стремиться максимумы значений в случае признания гипотезы о наличии прямой связи между доходами семьи и уровнем образования респондентов. А в таблице 5.10, которая по наполнению полностью повторяет предыдущую, выделены по два максимальных значения в каждом столбце.

Как видим из таблицы 5.9, гипотеза не работает, так как далеко не всегда максимальные значения стремятся к диагонали, а значит, переменные вовсе не обязательно связаны между собой, или эта связь не носит линейного характера.

В таблице 5.10 мы выделили по два максимальных значения в каждом столбце. Это дает нам возможность увидеть, как меняется значение одной переменной в случае изменения другой в большем доверительном интервале. При внимательном рассмотрении таблицы 5.10 видно, что некоторая закономерность в распределении данных все же прослеживается. В частности, мы видим, что доход семей большинства респондентов с начальным и непол-

ным средним образованием не превышает 10 тысяч рублей, респондентов, окончивших среднюю школу, как и тех, кто получил среднее специальное образование, находится в коридоре от 5 до 15 тысяч, с неоконченным высшим образованием – от 10 до 20 тысяч рублей. Доход семей респондентов с высшим образованием имеет больший коридор колебаний – от 5 до 25 тысяч рублей. Таким образом, мы можем говорить, что есть некоторая тенденция в повышении уровня доходов семей с повышением уровня образования респондентов, хотя, и это очевидно, тенденция зафиксирована весьма условно.

Таблица 5.9

Таблица сопряженности переменных «ДОХОД СЕМЬИ (Категоризовано)» и «ОБРАЗОВАНИЕ»

		S3. ОБРАЗОВАНИЕ								Итого
		Начальн.	Н/ср. шк. (9 кл.)	ПТУ на базе н/ср. школы	Ср. шк. (11 кл.)	ПТУ на базе ср. шк.	Ср. спец.	Н/высшее	Высшее	
D9. ДОХОД СЕМЬИ (Категоризовано)	< 5000	50,0%	33,1%	13,1%	19,4%	11,0%	13,1%	10,9%	6,4%	16,8%
	5000–9999	30,8%	35,5%	27,3%	20,2%	31,4%	19,5%	14,5%	18,7%	23,3%
	10000–14999	15,4%	15,1%	22,2%	21,0%	22,0%	21,3%	18,2%	16,0%	19,4%
	15000–19999		7,8%	12,1%	14,9%	13,6%	17,1%	16,4%	16,0%	14,0%
	20000–24999	1,9%	3,0%	11,1%	10,9%	10,2%	13,1%	14,5%	16,9%	11,3%
	25000–29999		1,2%	5,1%	4,8%	4,2%	4,3%	10,9%	6,4%	4,5%
	30000–34999	1,9%	3,0%	5,1%	3,2%	3,4%	5,1%	5,5%	7,8%	4,7%
	35000+		1,2%	4,0%	5,6%	4,2%	6,7%	9,1%	11,9%	6,1%
Итого		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Проверить тесноту связи позволяет вычисление коэффициента корреляции. Коэффициент корреляции принимает значение от –1 до +1, при этом чем ближе его значение к единице, тем более сильная связь существует между изучаемыми признаками. Знак «+» (плюс) или «–» (минус) указывает на направленность взаимосвязи. Прямая (положительная) взаимосвязь указывает на то, что с увеличением значения независимой переменной увеличивается значение зависимой, обратная (отрицательная) зависимость указывает на то, что при увеличении значения независимой переменной снижается значение зависимой. Важно знать, что значение коэффициента корреляции не зависит от масштаба или единицы измерения. Поэтому коэффициент корреляции очень удобен для понимания и объяснения взаимозависимостей различных по содержанию переменных (чего нельзя сказать о многих других коэффициентах, рассмотренных ранее).

Таблица 5.10

Таблица сопряженности D9. ДОХОД СЕМЬИ (Категоризовано) * S3. ОБРАЗОВАНИЕ

		S3. ОБРАЗОВАНИЕ								Итого
		Начальн.	Н/ср. шк. (9 кл.)	ПТУ на базе н/ср. школы	Ср. шк. (11 кл.)	ПТУ на базе ср. шк.	Ср. спец.	Н/выс-шее	Высшее	
D9. ДОХОД СЕМЬИ (Категоризовано)	< 5000	50,0%	33,1%	13,1%	19,4%	11,0%	13,1%	10,9%	6,4%	16,8%
	5000–9999	30,8%	35,5%	27,3%	20,2%	31,4%	19,5%	14,5%	18,7%	23,3%
	10000–14999	15,4%	15,1%	22,2%	21,0%	22,0%	21,3%	18,2%	16,0%	19,4%
	15000–19999		7,8%	12,1%	14,9%	13,6%	17,1%	16,4%	16,0%	14,0%
	20000–24999	1,9%	3,0%	11,1%	10,9%	10,2%	13,1%	14,5%	16,9%	11,3%
	25000–29999		1,2%	5,1%	4,8%	4,2%	4,3%	10,9%	6,4%	4,5%
	30000–34999	1,9%	3,0%	5,1%	3,2%	3,4%	5,1%	5,5%	7,8%	4,7%
	35000+		1,2%	4,0%	5,6%	4,2%	6,7%	9,1%	11,9%	6,1%
Итого		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Значение коэффициентов корреляции следует интерпретировать следующим образом:

$r < 0,2$	очень слабая корреляция
$0,2 \leq r < 0,5$	слабая корреляция
$0,5 \leq r < 0,7$	средняя корреляция
$0,8 \leq r < 0,9$	сильная корреляция
$r \geq 0,9$	очень сильная корреляция

В SPSS заложена возможность вычисления множества существующих коэффициентов корреляции. Среди них как самый простой и один из часто применяемых – коэффициент линейной корреляции Пирсона, так и разновидности коэффициентов ранговой корреляции Спирмена и Кендалла.

Один из часто используемых коэффициентов – *коэффициент корреляции Пирсона r* – называется также линейной корреляцией, так как измеряет степень линейных связей между переменными. Можно сказать, что корреляция определяет степень, с которой значения двух переменных пропорциональны друг другу. Пропорциональность означает простую линейную зависимость. Применение коэффициента линейной корреляции Пирсона предполагает, что рассматриваемые переменные измерены, как правило, в интервальной шкале, то есть являются количественными пере-

менными. Это, безусловно, несколько ограничивает возможности применения данного коэффициента корреляции в решении прикладных исследовательских задач.

При изучении социальных явлений социолог чаще сталкивается с качественными признаками, чем с количественными. Математико-статистическое изучение качественных признаков требует от исследователя обязательной работы по переводу качественной информации в количественные показатели. Одним из распространенных способов такого перевода данных является ранжирование, то есть расположение признака по убыванию или возрастанию частот. Для изучения взаимосвязи между ранжированными рядами в статистике и социологии применяются коэффициенты ранговой корреляции Спирмена и Кендалла. Эти коэффициенты используются и в том случае, если исследуемые признаки имеют ассиметричное распределение или если характер распределения не известен. При этом коэффициент Кендалла считается более взвешенным и математически обоснованным.

Коэффициент *ранговой корреляции Спирмена* является по сути аналогом коэффициента Пирсона для ранжированных рядов данных, а *коэффициент Кендалла* учитывает значения зависимых и независимых переменных, поэтому дает более осторожный результат при вычислениях. Вообще при вычислении коэффициентов ранговой корреляции следует помнить, что, в отличие от коэффициента Пирсона, ранговая корреляция фиксирует тесноту взаимосвязи между ранжированными рядами значений, а не между самими переменными.

Рассмотрев в общих чертах особенности корреляционных коэффициентов и условия их применения к социологическим данным, объясним, как можно вычислить значение тех или иных коэффициентов корреляции в SPSS на примерах.

Для начала попробуем изучить линейную зависимость в интервальной и номинальной шкалах с применением *коэффициента корреляции Пирсона r*. Этот коэффициент вычисляется в SPSS с помощью последовательности функций «Анализ» – «Описательные статистики» – «Таблицы сопряженности» – «Статистики». Для вычисления статистики Пирсона заменим использовавшиеся ранее ранжированные переменные – количественными (возраст респондента и количество туристических поездок по России за последние три года). Далее в окне «Статистики» поставим галочку напротив значения «Корреляции», затем нажмем кнопки «Продолжить» и «ОК» (см. рис. 5.7).

В результате этих действий получим таблицу 5.11. Как видим, в SPSS коэффициенты корреляции Пирсона и Спирмена выводятся автоматически в одной таблице, и исследователю просто нужно понимать, какой из коэффициентов следует использовать в тех или иных случаях.

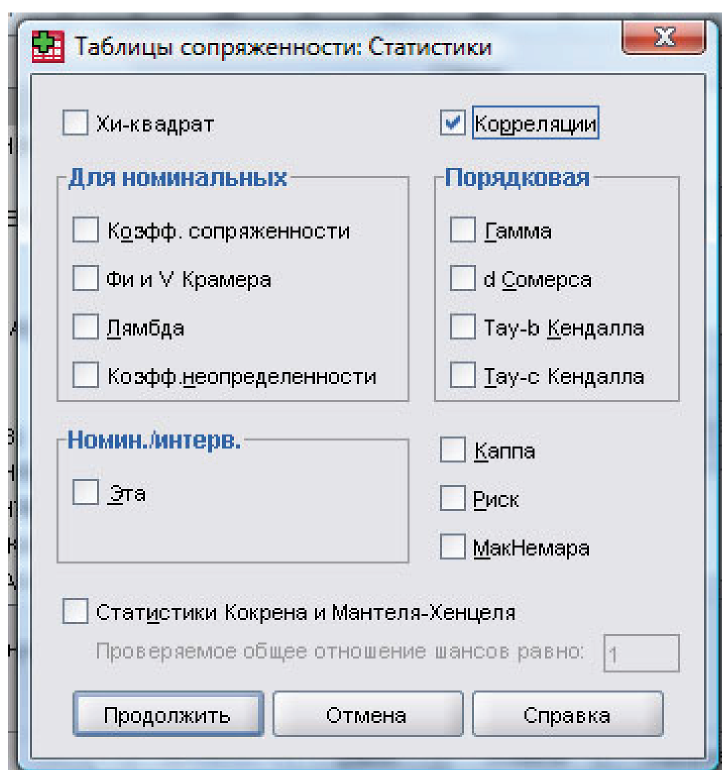


Рис. 5.7. Диалоговое окно «Таблицы сопряженности: Статистики»

Таблица 5.11

Симметричные меры

		Значение	Асимптотическая стандартная ошибка[a]	Прибл. T[b]	Прибл. значимость
Интервальная по интервальной	R Пирсона	,263	,021	10,882	,000[c]
Порядковая по порядковой	Корреляция Спирмена	,238	,022	9,783	,000[c]
Кол-во валидных наблюдений		1601			

[a] Не подразумевая истинность нулевой гипотезы.

[b] Используется асимптотическая стандартная ошибка в предположении истинности нулевой гипотезы.

[c] На основании нормальной аппроксимации.

В нашем примере коэффициент линейной корреляции Пирсона равен 0,263. Оценка доверительного интервала корреляции, основанная на асимптоматической стандартной ошибке и значении аппроксимации, приведенных в этой же таблице, показывает, что теснота связи двух переменных находится в пределе от 0,263 – 10,882 × 0,021 до 0,263 + 10,882 × 0,021 (то есть от 0,034 до 0,492). В любом случае это лишь слабая или очень слабая корреляция. Следовательно, на основе вычисления коэффициента корреляции в данном случае мы можем однозначно констатировать слабую взаимосвязь

между количеством туристических поездок по стране за последние годы и возрастом респондентов.

В том же диалоговом окне «Таблицы сопряженности: Статистики» можно выбрать и вычисление коэффициентов ранговой корреляции Кендалла, правда, для этого уже нужно сменить переменные с количественных на качественные – ранжированные. Проделаем это на примере, изучив наличие взаимосвязи между собственными оценками дохода семьи и уверенностью в завтрашнем дне. Первая переменная «Уверенность в завтрашнем дне» имеет позиции от 1 – «Определенно уверен» до 4 – «Определенно не уверен», вторая переменная «Оценка доходов семьи» со значениями от 1 – «Мы едва сводим концы с концами» до 5 – «Мы можем позволить себе достаточно дорогостоящие вещи». Обе шкалы порядковые, поэтому в данном случае нас интересует лишь коэффициенты корреляции Спирмена и Кендалла.

Приведенные в таблице 5.12 результаты расчетов корреляционных значений указывают на наличие хотя и слабой (0,263), но отрицательной корреляции. Это значит, что при возрастании значений одной переменной значения другой, как правило, снижаются. Напоминаем, что коэффициенты ранговой корреляции выявляют тесноту связи между ранжированными рядами данных, а не между самими переменными, то есть если их ранжировать иначе, то и значение коэффициентов может измениться.

Таблица 5.12

Симметричные меры

		Значение	Асимптотическая стдандартная ошибка[a]	Прибл. T[b]	Прибл. значимость
Порядковая по порядковой	Тау-b Кендалла	–,263	,020	–12,849	,000
	Тау-c Кендалла	–,242	,019	–12,849	,000
	Корреляция Спирмена	–,304	,023	–12,474	,000[c]
Интервальная по интервальной	R Пирсона	–,299	,023	–12,207	,000[c]
Кол-во валидных наблюдений		1525			

[a] Не подразумевая истинность нулевой гипотезы.

[b] Используется асимптотическая стандартная ошибка в предположении истинности нулевой гипотезы.

[c] На основании нормальной аппроксимации.

В содержательном плане для нашего примера вычисления коэффициентов ранговой корреляции означает, что при лучшей самооценке доходов семьи респонденты, как правило, лучше оценивают свою уверенность в завтрашнем дне, и наоборот. Хотя в данном случае значение коэффициента

невелико, но все же следует говорить об определенной взаимозависимости между этими переменными.

Важно отметить, что в обыденной жизни при изучении различных социальных явлений социологу бывает сложно найти строгие математические закономерности. Социальные явления (за небольшим исключением) не укладываются в линейные модели при интерпретации, поэтому социологу важно уметь расчленить сложное явление на ряд составляющих, которые можно изучить статистически с тем, чтобы на основе совокупности математических показателей судить более достоверно об изучаемом явлении в целом. В этом социологу существенно помогают группировочные таблицы, рассмотренные в данной лекции.

Вопросы и задания

1. Для чего нужны таблицы сопряженности?
2. Опишите процедуру построения таблицы сопряженности в SPSS.
3. В чем разница между столбцовой и строковой переменной? Что характеризуют те и другие?
4. Что такое критерий хи-квадрат? Для чего он применяется? Каковы условия правомерности его применения к социологическим данным?
5. Опишите основные принципы интерпретации коэффициентов корреляции.
6. Существует ли различие между коэффициентами корреляции Пирсона, Спирмена и Кендалла?

Список литературы

1. Аптон, Г. Анализ таблиц сопряженности/Г. Аптон. — М.: Финансы и статистика, 1982. — 143 с.

2. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей/А. Бююль, П. Цёфель. — СПб.: ДиаСофтЮП, 2005. — 608 с.

3. Езекиел, М. Методы анализа корреляций и регрессий/М. Езекиел, К. Фокс. — М.: Статистика, 1966. — 361 с.

4. Елисеева, И. И. Группировка, корреляция, распознавание образов/И. И. Елисеева, В. О. Рукавишников. — М.: Статистика, 1977. — 144 с.

5. Крыштановский, А. О. Анализ социологических данных/А. О. Крыштановский. — М.: Изд-во «ГУ ВШЭ», 2007. — 281 с.

Лекция 6. Регрессионный анализ

В настоящем разделе рассмотрим процедуру регрессионного анализа. Регрессионный анализ – это статистический метод, который позволяет исследовать зависимость между двумя переменными, одна из которых является зависимой, а остальные – независимыми[12].

Существует несколько видов регрессионного анализа, часть из которых, как видно из рисунка 6.1, реализованы в пакете SPSS. В рамках настоящей лекций остановимся детально на линейной регрессии и, поняв ее суть, кратко охарактеризуем остальные виды регрессионного анализа.

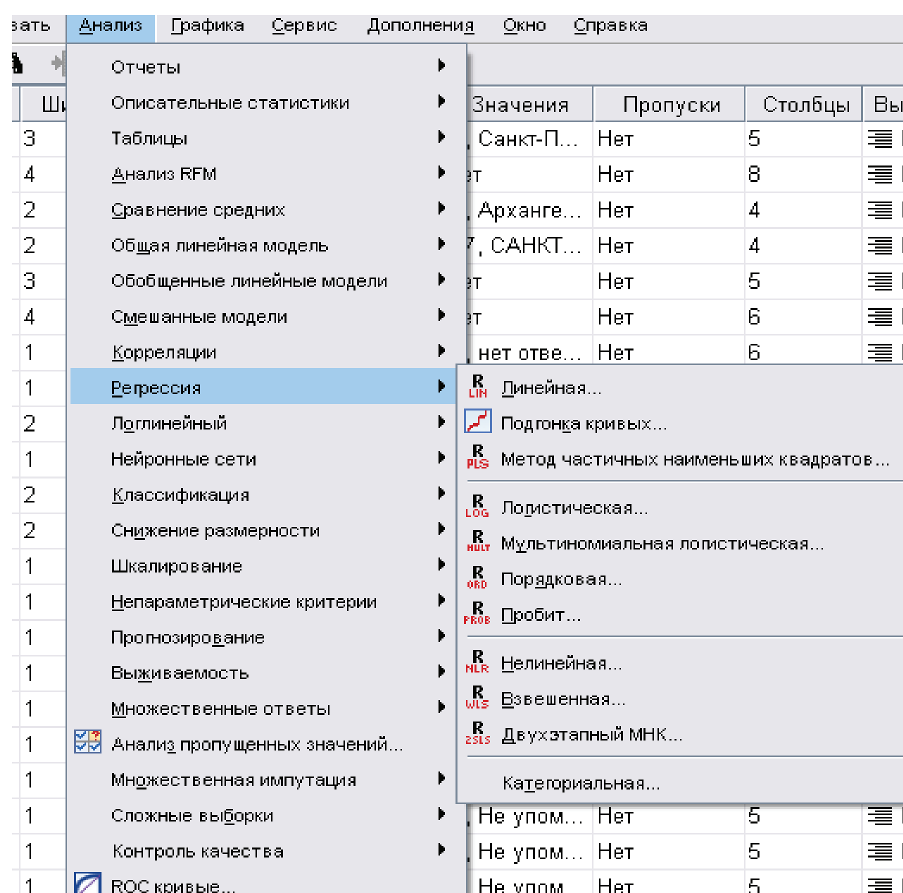


Рис. 6.1. Раздел «Регрессия» в меню «Анализ»

Линейная регрессия является наиболее простым методом оценки коэффициентов линейного уравнения, содержащего одну или несколько независимых переменных, которые позволяют наилучшим образом предсказать значение зависимой переменной. Например, можно попытаться предсказать объем годовых продаж для сотрудника отдела продаж (зависимая

---

[12] Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

переменная) по таким независимым переменным, как возраст, образование и стаж работы.

Прежде чем переходить к рассмотрению процедуры «Линейная регрессия» в SPSS, сделаем некоторые пояснения с целью облегчения восприятия последующей информации. Поговорим о линейности. Что означает понятие «линейность»? Что такое «линейная связь», «линейное уравнение»? Ответим на эти вопросы.

Понятие «линия» происходит от латинского Linea – букв. льняная нить; линия, черта. Линейность – свойство систем или процессов, заключающееся в наличии линейной (прямой) зависимости одних факторов от других. Можно привести много примеров линейной взаимосвязи: это зависимость времени, потраченного на дорогу, от скорости передвижения; зависимость массы дерева от его возраста, зависимость уровня заработной платы человека от стажа работы и т.д.

Необходимо отметить, что по мере усложнения уровня организации систем, наличие в них линейностей встречается все реже и реже. Так, на уровне элементарных частиц, базовых физических законов линейные взаимосвязи очень распространены. Достаточно часто линейные закономерности встречаются и в экономике. Реже в маркетинге. Но в такой сложной системе, как общество, все взаимосвязи носят в основном нелинейный характер. Поэму в процессе анализа данных опросов общественного мнения метод линейной регрессии практически не используется. Тем не менее очень важно рассмотреть этот вид регрессии для того, чтобы иметь полное представление о регрессионном анализе в целом.

Рассмотрим реализацию метода линейной регрессии на примере, который мы взяли из учебника А.О. Крыштановского «Анализ социологических данных с помощью пакета SPSS» [13], немного изменив его для упрощения расчетов.

Итак, выяснение причин хорошей или плохой успеваемости студентов является, несомненно, сложной задачей. Социологические теории, да и просто здравый смысл подсказывают нам, что среди факторов, влияющих на успеваемость, должны присутствовать:

- уровень подготовки студента;
- активность посещения занятий;
- активность самостоятельной работы;
- способности студента.

Очевидно, что этот список неполон и может быть расширен за счет других характеристик, однако ограничимся пока только этими.

[13] Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS // А.О. Крыштановский – М.: Изд.дом ГУ ВШЭ, 2006.

Представим схему влияния различных показателей на успеваемость в виде рисунка (рис. 6.2).



Рис. 6.2. Модель «Успеваемость студента»

Данный рисунок можно рассматривать как модель успеваемости, или некоторую схему, которая позволяет систематизировать наши взгляды на изучаемое явление. Анализируя эмпирическое данные, можно попытаться проверить, насколько наша модель соответствует тем реальным процессам, которые управляют успеваемостью и данные о которых можно собрать с помощью социологических методов.

Исходя из объема пройденного материала в рамках настоящего курса, можно констатировать, что в нашем распоряжении есть только инструменты проверки парных взаимосвязей между переменными – коэффициенты сопряженности и корреляции. При этом сами коэффициенты фактически фиксируют не то, насколько сильно взаимосвязаны два показателя между собой, а то, насколько тесно они взаимосвязаны.

Теснота взаимосвязи является, несомненно, важной характеристикой, но на практике интереснее сила связи. Так, мы знаем – если солить еду, она становится солонее. Другими словами, эти характеристики взаимосвязаны, и, по всей видимости, достаточно тесно. Однако крайне важно знать и то, насколько становится солонее блюдо при добавлении определенного количества соли. Зависит это и от характеристик соли, и от особенностей используемых продуктов, и от специфики процесса приготовления. Согласитесь, без этого знания вкусного блюда не приготовишь.

В модели, представленной на рис. 6.2, для нас принципиально важно не только наличие обозначенных стрелок. Чтобы модель давала нам полезную информацию, которую можно использовать на практике, необходимо иметь представление о силе соответствующих связей, т.е. понимать, какие из показателей влияют на успеваемость сильнее, а какие слабее, а также на-

сколько велико совокупное влияние на успеваемость четырех выделенных факторов.

Решение поставленной задачи начнем с упрощения модели, представленной на рис. 6.2, к модели, представленной на рис. 6.3.



Рис. 6.3. Упрощенная модель «Успеваемость студента»

Отличие модели с рисунка 6.3 от модели с рисунка 6.2 состоит в том, что мы фокусируемся только на одной причине успеваемости студента – уровне предварительной подготовки, а все остальные факторы включили в «Другие факторы». Социологический смысл данной модели представляется вполне естественным: успеваемость студента зависит от уровня его предварительной подготовки. Разумеется, успеваемость определяется не только этим. Имеется еще множество других факторов, влияющих на успеваемость. Смысл построения модели математической зависимости состоит в выяснении того, каким образом на успеваемость влияет именно уровень предварительной подготовки, каково направление и сила этого влияния.

Если о направлении воздействия можно сделать, как представляется, вполне обоснованное предположение «чем выше уровень предварительной подготовки, тем выше успеваемость», то сформулировать предположения о силе такого воздействия довольно сложно. Попытаемся с помощью анализа данных, содержащих сведения об успеваемости студентов и уровне их предварительной подготовки, найти точные ответы на поставленные вопросы.

Формально предложенную модель зависимости можно записать в виде следующей математической зависимости:

$$y=f(x) + u \tag{6.1}$$

где:

y – показатель «Успеваемость студента»;

x – показатель «Уровень предварительной подготовки»;

f – функция описывающая силу и форму влияния x на y;

u – все остальные факторы, влияющие на y.

Задачей построения модели рис 6.2. становится подбор функции, которая будет наилучшим образом описывать зависимость x от y. Рассмотрим решение этой задачи на примере.

В нашем распоряжении есть данные, в которых в качестве показателя «Уровень предварительной подготовки» выступает суммарный балл, полу-

ченный студентом на вступительных экзаменах в вуз, в качестве показателя «Успеваемость» – суммарный балл студента за первый семестр обучения в вузе (табл. 6.1)[14].

Таблица 6.1

Оценки студентов при поступлении в вуз и по итогам первого семестра обучения

№ студента	Суммарный балл на вступительных экзаменах (x)	Суммарный балл по итогам первого семестра Обучения (y)
1	32	117,4
2	26	106,7
3	27	120,0
4	27	97,3
5	26	108,0
6	25	124,0
7	25	121,4
8	28	106,7
9	29	105,3
10	27	96,0
11	26	94,7
12	26	89,4
13	25	113,4
14	26	113,3
15	24	93,3
16	25	118,7
17	25	88,0
IS	28	100,0
19	14	78,7
20	18	102,7

[14] Были взяты оценки абитуриентов на вступительных экзаменах в 2002 г. на факультет социологии ГУ ВШЭ. Вступительные испытания проводились по четырем дисциплинам: математика, обществознание, иностранный язык, русский язык. Оценки по первым трем дисциплинам выставлялись по 10-балльной системе, по русскому языку – по 5-балльной системе.

Все оценки за обучение в ГУ ВШЭ выставляются по 10-балльной системе, независимо от формы контроля (как за экзамены, так и за зачеты). При вычислении суммарного балла за семестр оценка по каждому предмету учитывается с определенным весом, который отражает объем часов по данному предмету. Так. если на предмет отводится, скажем, 50 часов, вес его оценки – 1, а если 100 часов, то вес оценки уже 2. Максимально возможная сумма баллов, которые мог набрать студент I курса в первом семестре 2002/03 учебного года, – 146,7.

Коэффициент корреляции Пирсона между двумя анализируемыми показателями составляет 0,43 и значим на уровне $a = 0,06$ (это было выявлено в процессе анализа таблиц сопряженности рассматриваемых переменных). Следовательно, у нас есть неплохие основания заключить, что модель, приведенная на рис. 6.3, отражает реально существующие закономерности. Представим данные таблицы 6.1 в виде диаграммы рассеяния (рис. 6.4).

Рисунок 6.4 показывает, что есть определенная зависимость между x и y – с ростом значений показателя «Уровень предварительной подготовки» наблюдается тенденция возрастания показателя «Успеваемость». Какова форма этой зависимости, или каков вид функции f в выражении (6.1)? Начнем поиск этой функции с самого простого и удобного класса функций – с линейных функций.

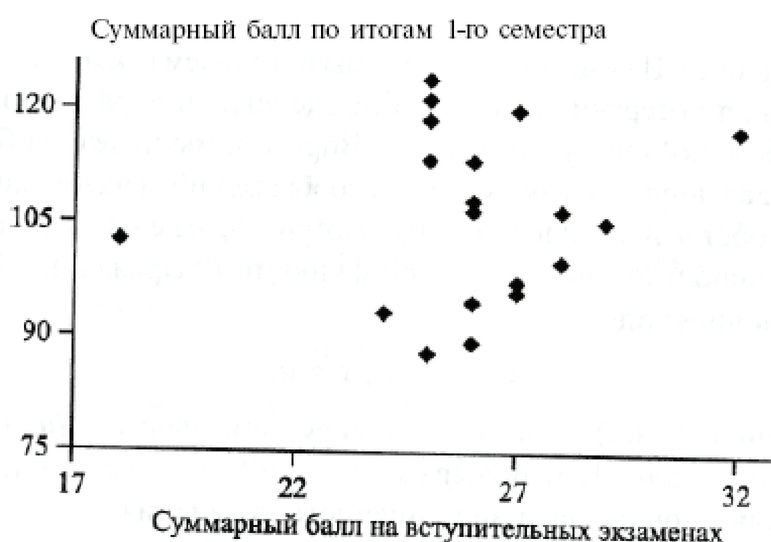


Рис. 6.4. Оценки студентов при поступлении в вуз и по итогам первого семестра обучения

Почему именно с линейных? Ведь диаграмма (см. рис. 6.4) показывает нам лишь то, что это должна быть какая-то возрастающая функция, а в этом качестве могут выступать и показательная функция, и логарифм, да и вообше бесконечное число самых разных функций. Причем видно, что какую бы функцию мы ни взяли, она не будет точно проходить через все точки.

Однако этого и не требуется. Ведь в выражении (6.1) значения y описываются не как f(x), а как сумма f(x) и u. Таким образом, можно сказать, что несовпадения положения точек с графиком некоторой функции f объясняются наличием именно добавки, которая, напомним, является «другими» факторами, влияющими на успеваемость студентов.

Данные соображения, к сожалению, не объясняют, почему мы решили рассматривать именно линейные функции. Объяснение этому лежит совсем

в другой плоскости – на самом деле линейные функции проще и удобнее. В некотором смысле мы поступаем как герой анекдота, который ищет потерянные часы не там, где он их потерял, а под фонарным столбом, поскольку там светлее. Впрочем, мы не всегда будем решать поставленную задачу исходя из соображений максимизации простоты и удобства, и далее рассмотрим другие виды функций.

При использовании линейной функции *f* выражение (6.1) примет следующий вид:

$$Y = b_0 + b_1x + u \quad (6.2)$$

Это уравнение называется уравнением простой (или парной) линейной регрессии. Здесь b0 и b1 – константы, которые и определяют конкретный вид линейного уравнения.

Представим, как будет выглядеть рис. 6.4, если на нем изобразить линейную функцию (6.2) (рис. 6.5).

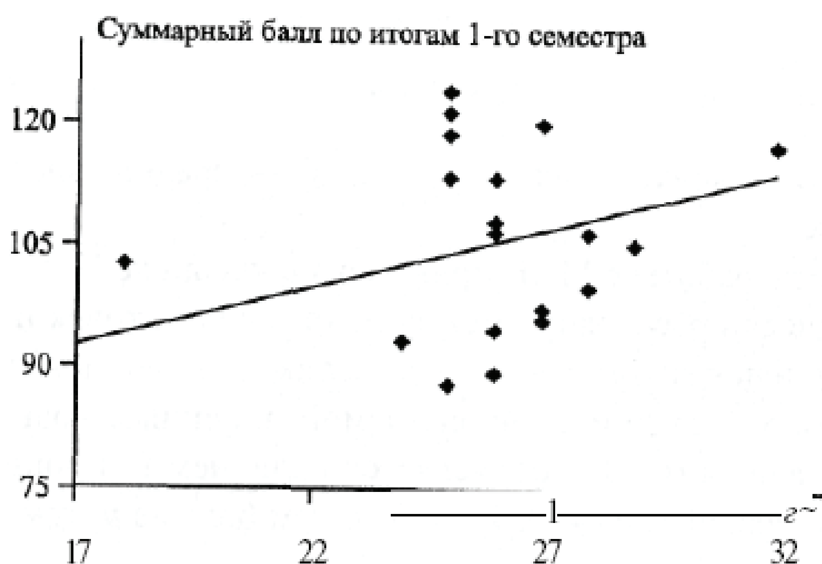


Рис. 6.5. Оценки студентов при поступлении в вуз и по итогам первого семестра обучения с изображением линейной функции

Из каких соображений мы исходили, строя прямую на рис. 6.5? Иными словами, как мы определили параметры b0 и b1, которые и дали нам именно такую прямую? Логика вычисления параметров прямой достаточно проста. Прямая должна лежать максимально близко ко всем точкам графика, т.е. сумма расстояний от всех точек до искомой прямой была бы наименьшей. Подробнее это показано на рис. 6.6.

Оставим для наглядности на графике четыре точки, а остальные сделаем невидимыми. Стрелки E1, E2, E3, E4 – это расстояния до регрессион-

ной прямой для точек 1, 2, 3, 4 соответственно. Один из способов вычисления параметров b0 и b1 регрессионного уравнения состоит в минимизации суммы S (6.3).

$$S = E1^2 + E2^2 + E3^2 + E4^2 \qquad (6.3)$$

Иначе говоря, мы стараемся сделать минимальной не сумму расстояний от точек до прямой, а сумму квадратов расстояний.

Фактически расстояния между положениями точек и регрессионной прямой показывают, насколько велико отличие между моделью зависимости *y* от *x*, описываемой линейным уравнением, и реальными данными. Это объясняется наличием величины u в регрессионном уравнении (6.2). Ясно, что чем больше u, тем хуже описывает линейная функция реальные данные.

Степень расхождения реальных данных *y* и *x*, вычисленных с помощью найденной функции, в регрессионном анализе называют остатками. На рис. 6.5 расстояния E1, E2, E3 и E4 и есть остатки.

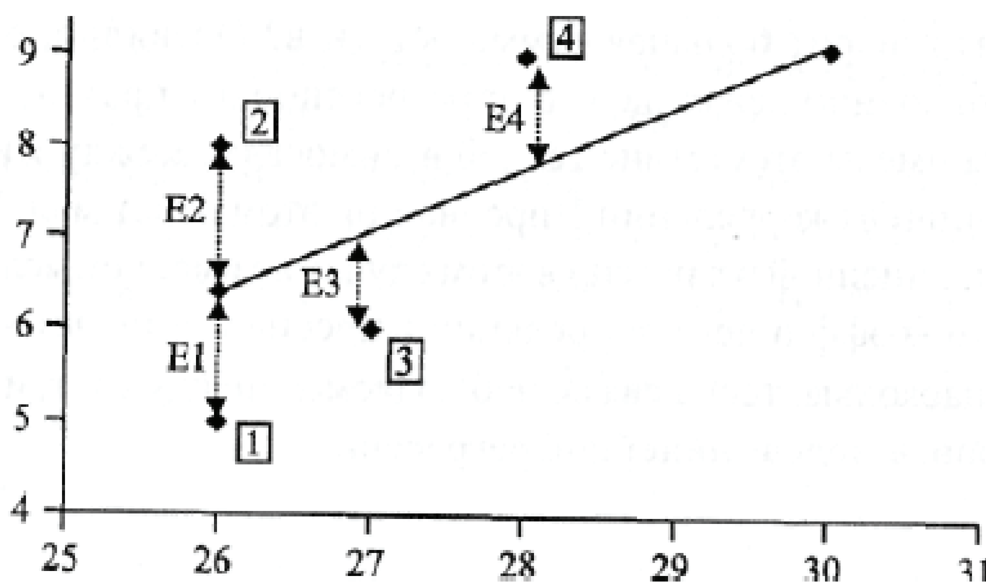


Рис. 6.6. Оценки студентов при поступлении в вуз и по итогам первого семестра обучения. Пример с четырьмя наблюдениями

О чем говорит большая сумма остатков? Очевидно, о том, что данные в основном лежат далеко от регрессионной прямой. Следовательно, мы имеем отсутствие тесной взаимосвязи между *x* и *y*. Ясно, что коэффициент корреляции Пирсона при этом будет мал. Построение модели линейной регрессии в этом случае не имеет смысла. Можно сказать, что коэффициент корреляции Пирсона выступает индикатором того, насколько тесна связь, наблюдаемая между *x* и *y*, и имеет ли смысл строить модель линейной регрессии.

Метод решения задачи вычисления параметров регрессии путем минимизации выражения (6.3) называется методом наименьших квадратов (МНК). Оказывается, что S минимальна при значениях b0 и b1, представленных в формулах (6.4), (6.5); b1 иначе называют «бета-коэффициентом», который рассчитывается как отношение ковариации двух переменных к дисперсии второй переменной (6.4).

$$b_1 = \frac{\operatorname{cov}(x,y)}{D_x} \quad (6.4)$$

$$b_0 = \overline{y} - b_1\overline{x} \quad (6.5)$$

Где:

$\operatorname{cov}(x,y)$ – ковариация *x* и *y*;

D_x – дисперсия переменной *x*;

$\overline{x}$ и $\overline{y}$ – средние значения этих переменных.

Ковариация несет тот же смысл, что и коэффициент корелляции, то есть показывает, есть ли взаимосвязь между двумя случайными величинами. Рассчитывается по формуле

$$\operatorname{cov}(x,y) = 1/n\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \text{ или } \operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}, \quad (6.6)$$

где

$\overline{x}$ и $\overline{y}$ – средние значения этих переменных;

x_i и y_i – показатели линейной функции (в нашем случае показатели «Успеваемость студента» и «Уровень предварительной подготовки»);

n – общее количество показателей линейной функции (в нашем случае это значение равно 20 – т.е. общее количество студентов, на основании успеваемости которых строится модель). Для наглядности рассчитаем «вручную» линейную функцию, используя данные, приведенные выше. Для начала найдем средние значения для успеваемости студентов и уровня их предварительной подготовки.

$\overline{x}$ = (32 + 26 + 27 + 27 + 26 + 25 + 25 + 28 + 29 + 27 + 26 + 26 + 25 + 26 + 24 + 25 + 25 + 28 + 14 + 18) / 20 = 25,45

$\overline{y}$ = (117,4 + 106,7 + 120,0 + 97,3 + 108,0 + 124,0 + 121,4 + 106,7 + 105,3 + 96,0 + 94,7 + 89,4 + 113,4 + 113,3 + 93,3 + 118,7 + 88,0 + 100,0 + 78,7 + 102,7) = 104,75

Вычислим ковариацию

$\text{cov}(x,y)$ = [(32 – 25,45) * (117,4 – 104,75) + (26 – 25,45) * (106,7–104,75) + …и т.д.] / 20 = 381,15 / 20 = 19,0575

Вычислим дисперсию $D_{x.}$ Существует множество видов расчета дисперсии. Не вдаваясь в детали различных способов, отметим, что программа SPSS рассчитывает дисперсию формуле (6.7)

$$D_x = \frac{\sum (x-\overline{x})^2}{(n-1)} \qquad (6.7)$$

В нашем случае

$D_{x.}$= [(32 – 25,45)^2 + (26 – 25,45)^2+…и т.д.] / 20 – 1 = 266,95 / 19 = 14,1

Итак, получаем бета-коэффициент нашей регрессионной модели:

$$b_1 = \frac{\text{cov}(x,y)}{D_x} = \frac{19,0575}{14,05} \approx 1,4$$

Для получения второго коэффициента подставляем в формулу (6.5) значение бета-коэффициента, получаем:

$$b_0 = \overline{y} - b_1\overline{x} = 104,75 - 1,4 * 25,45 = 104,75 - 35,63 = 69,12$$

Итоговая формула регрессионной модели в нашем примере будет выглядеть следующим образом:

$$y = 69 + 1,4x \qquad (6.8)$$

Интерпретация коэффициентов регрессии

Коэффициент b_0 показывает, в какой точке регрессионная прямая пересечет ось *y*. Интерпретировать этот показатель достаточно просто: какую успеваемость по итогам первого семестра будут иметь студенты, которые набрали на вступительных экзаменах 0 баллов. Они будут иметь успеваемость 69 баллов. Очевидно, в рамках данного примера такая ситуация бессмысленна, однако во многих случаях b_0 несет полезную информацию.

Смысл коэффициента b_1 интереснее. Он показывает, на сколько баллов возрастает средняя успеваемость студента в первом семестре при увеличении на единицу балла на вступительных экзаменах в вуз. Таким образом, мы видим, что увеличение суммарной оценки на вступительных экзаменах на 1 балл дает улучшение успеваемости студента в первом семестре на 1,4 балла. На самом деле коэффициент b_1 есть не что иное, как тангенс угла наклона регрессионной прямой, и, следовательно, именно он демонстрирует силу связи между *x* и *y*.

Практическая часть

Все переменные для процедуры линейной регрессии (зависимые и независимые) должны быть количественными. Использование других типов переменных не допускается.

Для линейного регрессионного анализа в учебных целях возьмем переменные доход и количество часов в неделю, которое респондент занят на работе[15]. В нашем случае независимой переменной будет являться вопрос анкеты «Сколько часов в неделю Вы обычно заняты на работе?», зависимой – «Какова была сумма лично Ваших заработков, доходов от собственного бизнеса и т.п. за последний месяц (тыс. руб.)?». Гипотезой является то, что доход напрямую зависит от времени, проводимого на работе. Итак, проверим нашу гипотезу. Для этого вызываем диалоговое окно «Линейная регрессия». В разделы «Зависимые переменные» и «Независимые переменные» вводим соответствующие, указанные выше, переменные (рис. 6.7).

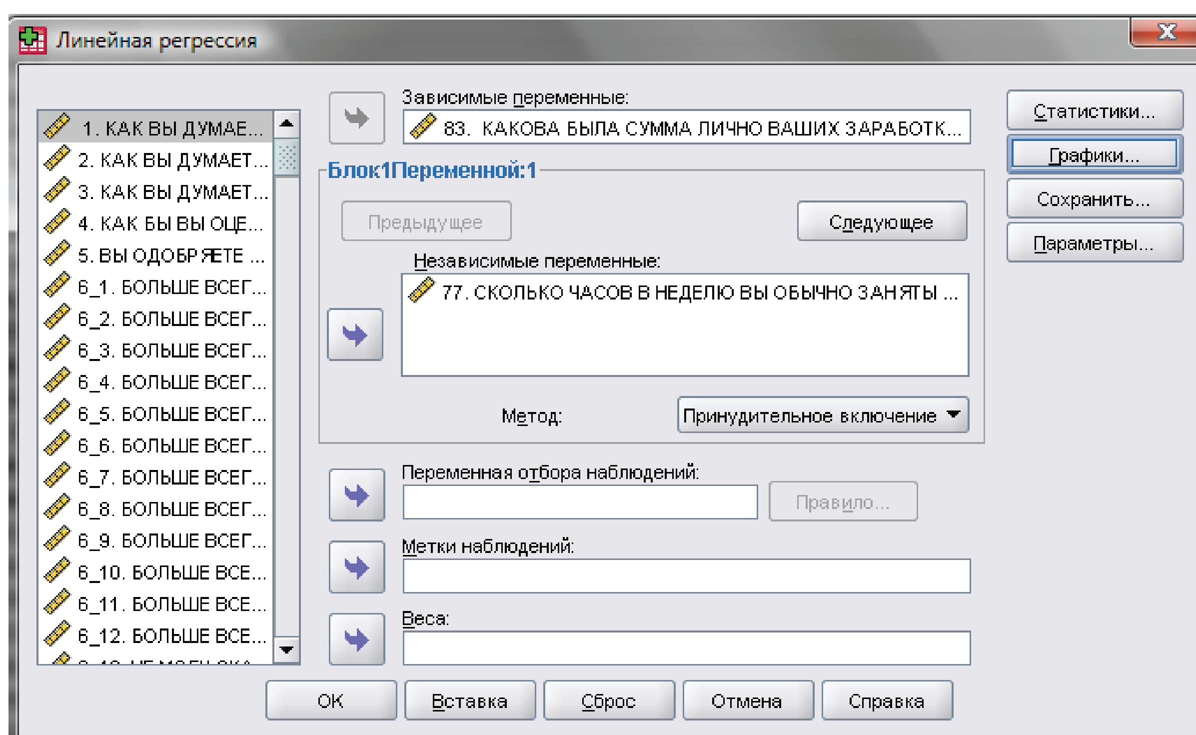


Рис. 6.7. Диалоговое окно «Линейная регрессия»

Если бы в исследовании была необходимость построения регрессионных моделей для отдельных групп респондентов, то в поле «переменная отбора наблюдений» диалогового окна «Линейная регрессия» необходимо

[15] Использовались данные Единого архива данных социологических исследований. Исследование «Факт», 1993 г., 8-я волна. Посвящено социальным проблемам, социальным страхам и тревогам, экологии. Выборка всероссийская многоступенчатая стратифицированная случайная. Исполнитель ВЦИОМ.

было бы указать переменную, по которой производится отбор респондентов в исследуемые группы. Например, если бы нужно было построить две регрессионные модели для мужчин и женщин, то из списка переменных в указанное поле следовало бы перенести метку переменной «Пол респондента».

В правом верхнем углу диалогового окна «Линейная регрессия» имеются четыре кнопки: «Статистики», «Графики», «Сохранить», и «Параметры». При нажатии на них открываются дополнительные диалоговые окна. Так, при нажатии кнопки «Статистики», на экране появляется соответствующее диалоговое окно (рис. 6.8).

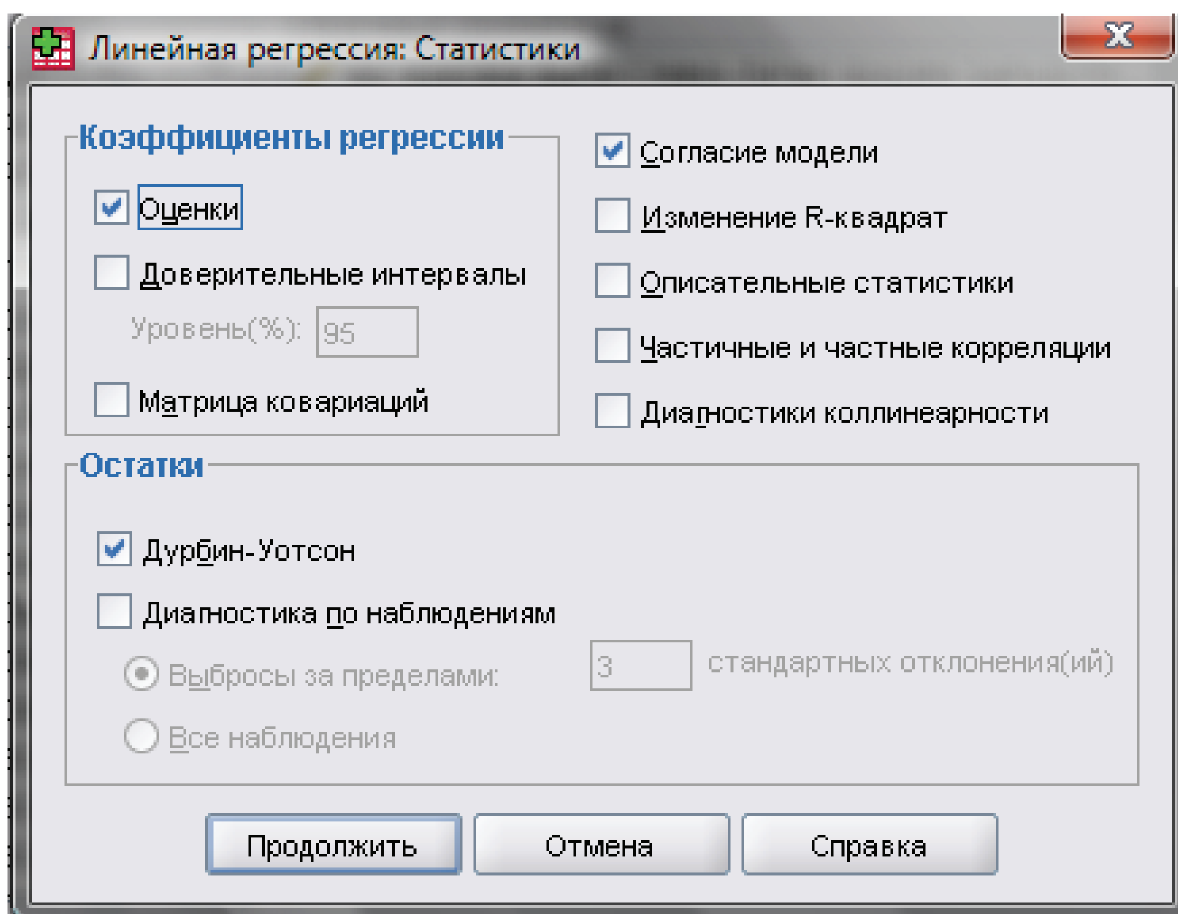


Рис. 6.8. Диалоговое окно «Линейная регрессия: Статистики»

Здесь можно выбрать различные статистические показатели. В нашем примере отметим следующие команды: в разделе «коэффициенты регрессии» выберем пункт «Оценки», в правой части диалогового окна – «Согласие модели», а в разделе «Остатки» – «Дурбин-Уотсон».

Пункт «Оценки» рассчитает коэффициент детерминации (R), R-квадрат и некоторые другие статистические показатели, которые нужны для оценки качества построенной модели. Тест Дурбина-Уотсона (Durbin – Watson Test) необходим для вычисления показателей, используемых для анализа остат-

ков[16]. А пункт «Согласие модели» позволяет проверить статистическую гипотезу[17].

Итак, при нажатии кнопки «Продолжить» окно закрывается, и мы возвращаемся в главное диалоговое окно «Линейная регрессия», где после нажатия кнопки «ОК» запускается процедура выполнения простого регрессионного анализа. В итоге в окне вывода программа выдает нам информацию в виде нескольких таблиц. Первая таблица дает общую информацию о модели, полученной в результате линейной регрессии.

В таблице 6.2 представлены показатели, оценивающие качество линейной модели, построенной в результате проведения регрессионного анализа (это результат выбора пункта «Оценки»).

Таблица 6.2

Качество линейной модели

Сводка для модели[b]					
Модель	Н	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки	Дурбин-Уотсон
1	,145[a]	,021	,020	37,9779	1,753
[a] Предикторы: (конст) 77. СКОЛЬКО ЧАСОВ В НЕДЕЛЮ ВЫ ОБЫЧНО ЗАНЯТЫ НА РАБОТЕ?					
[b] Зависимая переменная: 83. КАКОВА БЫЛА СУММА ЛИЧНО ВАШИХ ЗАРАБОТКОВ, ДОХОДОВ ОТ СОБСТВЕННОГО БИЗНЕСА И Т.П. ЗА ПОСЛЕДНИЙ МЕСЯЦ (ТЫС.РУБ.)?					

Рассмотрим основные показатели этой таблицы. Н – Коэффициент корреляции между наблюдаемыми и предсказанными значениями зависимой переменной; варьируется в пределах от 0 до 1. Малая величина показывает, что линейная зависимость между зависимой и независимой переменными очень низкая или ее нет совсем. В нашем случае значение 0,145 показывает, что зависимость очень низкая.

R-квадрат – мера линейной модели, которую иногда называют коэффициентом детерминации. Он также варьируется в пределах от 0 до 1. Малые значения свидетельствуют о том, что модель плохо вписывается в данные. В нашем примере значение коэффициента 0,021 означает, что построенная

[16] Остатки должны появляться случайно, т.е. не ситематически. Для проверки этого условия проводится тест Дурбина–Уотсона. В ходе проведения этого теста рассчитывается коэффициент, значение которого лежит в диапазоне от 0 до 4. Если значение коэффициента близко к среднему (т.е. к 2), это означает, что автокорреляция отсутствует, т.е. остатки появляются случайным образом.

[17] Проверить статистическую гипотезу – это значит проверить, согласуются ли данные, полученные из выборки с этой гипотезой. Проверка осуществляется с помощью статистического критерия. Статистический критерий – это случайная величина, закон распределения которой (вместе со значениями параметров) известен в случае, если принятая гипотеза справедлива. Этот критерий называют еще критерием согласия (имеется в виду согласие принятой гипотезы с результатами, полученными из выборки).

регрессионная модель описывает лишь 2,1% случаев, когда увеличение количества часов, проведенных на работе, влечет за собой увеличение доходов

Значение теста Дурбина–Уотсона составляет 1,753, что близко к 2. Это говорит об отсутствии систематических связей между остатками, т.е. между отклонениями наблюдаемых (эмпирических) значений от теоретически ожидаемых (расчетных).

Следующая таблица показывает результаты теста однофакторного дисперсионного анализа (которая также является результатом выбора пункта «Оценки»). Важнейшим ее показателем является последняя колонка, в которой отражено значение показателя «Статистическая значимость». Значение данного показателя, меньшее или равное 0,5, показывает, что регрессионная модель, построенная на основе данных респондентов, попавших в выборку, справедлива для всей генеральной совокупности в целом. В нашем случае это именно так.

Таблица 6.3

Результаты теста однофакторного дисперсионного анализа

Дисперсионный анализ[b]

Модель		Сумма квадратов	ст.св.	Средний квадрат	Щ	Знч.
1	Регрессия	30243,138	1	30243,138	20,968	,000[a]
	Остаток	1410586,190	978	1442,317		
	Всего	1440829,327	979			

[a] Предикторы: (конст) 77. СКОЛЬКО ЧАСОВ В НЕДЕЛЮ ВЫ ОБЫЧНО ЗАНЯТЫ НА РАБОТЕ?

[b] Зависимая переменная: 83. КАКОВА БЫЛА СУММА ЛИЧНО ВАШИХ ЗАРАБОТКОВ, ДОХОДОВ ОТ СОБСТВЕННОГО БИЗНЕСА И Т.П. ЗА ПОСЛЕДНИЙ МЕСЯЦ (ТЫС.РУБ.)?

Результаты регрессионного анализа, описывающие построенную регрессионную модель, представлены в табл. 6.4.

Таблица 6.4

Результаты регрессионного анализа

Коэффициенты

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	17,408	5,709		3,049	,002
	77. СКОЛЬКО ЧАСОВ В НЕДЕЛЮ ВЫ ОБЫЧНО ЗАНЯТЫ НА РАБОТЕ?	,638	,139	,145	4,579	,000

[a] Зависимая переменная: 83. КАКОВА БЫЛА СУММА ЛИЧНО ВАШИХ ЗАРАБОТКОВ, ДОХОДОВ ОТ СОБСТВЕННОГО БИЗНЕСА И Т.П. ЗА ПОСЛЕДНИЙ МЕСЯЦ (ТЫС.РУБ.)?

В столбце «B» таблицы представлены основные параметры полученной регрессионной модели. В нашем случае уравнение регрессии будет иметь вид: $y = 17,408 + 0,638c$. Это уравнение можно интерпретировать следующим образом: если $c = 0$, то есть респондент не тратит ни одного часа своего времени на работу (допустим, он безработный), то он будет получать при этом 17, 408 рублей (17,408 = 0,638×0). По мере увеличения значения «с», соответственно увеличивается и значение y.

В следующем столбце таблицы представлены стандартные ошибки. При доверительном интервале 95% каждый коэффициент может отклоняться от средней величины на $\pm 2 \times z$ (где z – стандартная ошибка). Например, общая сумма доходов респондента при нулевом количестве времени, проведенном на работе, может отклоняться от среднего значения (17,408 рублей) на ±2 × 5,709, то есть на ±11,418.

Значение коэффициента регрессии независимой переменной «77. Сколько часов в неделю Вы обычно заняты на работе?» составляет 0,638. Это означает, что увеличение времени, проведенного на работе, на 1 час в построенной модели влечет за собой увеличение дохода на 0,638 рублей.

Последняя таблица показывает нам статистику остатков (табл. 6.5)

Таблица 6.5

Статистики остатков проведенного регрессионного анализа

Статистики остатков[a]					
	Минимум	Максимум	Для среднего	Стд. Отклонение	M
Предсказанное значение	23,790	74,843	42,955	5,5580	980
Остаток	–61,8426	436,6443	,0000	37,9584	980
Стд. Предсказанное значение	–3,448	5,737	,000	1,000	980
Стд. Остаток	–1,628	11,497	,000	,999	980
[a] Зависимая переменная: 83. КАКОВА БЫЛА СУММА ЛИЧНО ВАШИХ ЗАРАБОТКОВ, ДОХОДОВ ОТ СОБСТВЕННОГО БИЗНЕСА И Т.П. ЗА ПОСЛЕДНИЙ МЕСЯЦ (ТЫС. РУБ.)?					

Здесь не будем подробно останавливаться на результатах приведенной таблицы, отметим лишь, что об остатках было сказано в теоретической части данной лекции.

Предсказанные значения, остатки и другие статистики, полезные для диагностики, можно сохранить. Выбор каждого из перечисленных ниже пунктов (рис 6.9) добавляет к активному файлу данных одну или несколько переменных.

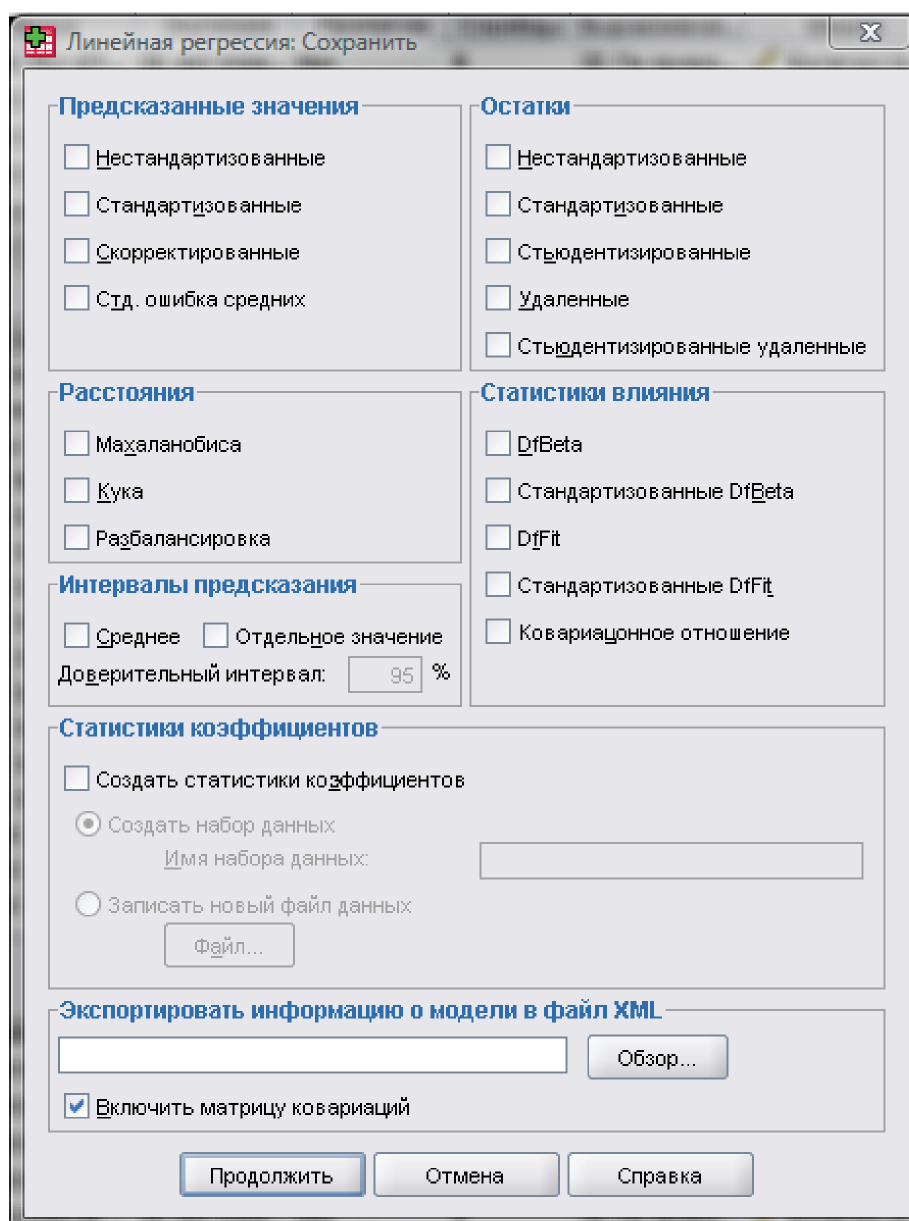


Рис. 6.9. Диалоговое окно «Линейная регрессия: Сохранить»

Графики процедуры «Линейная регрессия»

Графики могут помочь при проверке предположений о нормальности, линейности и равенстве дисперсий. Графики полезны также для выявления выбросов, необычных наблюдений и влияющих наблюдений. Сохраненные в качестве новых переменных предсказанные значения, остатки и другие диагностические величины становятся доступными в Редакторе данных. Их можно использовать в сочетании с независимыми переменными для построения графиков. В SPSS можно построить различные графики. *Диаграммы рассеяния* строятся для любой пары переменных из следующего списка: зависимая переменная, стандартизованные предсказанные значения, стандартизованные остатки, удаленные остатки, скорректированные предсказанные

значения, стьюдентизированные остатки, стьюдентизированные удаленные остатки. Для проверки линейности и равенства дисперсий строится график стандартизованных остатков против стандартизованных предсказанных значений. Для построения *графика линейной регрессии* нажимаем в правом верхнем углу диалогового окна «Линейная регрессия» кнопку «Графики» и в появившемся диалоговом окне выбираем графики, которые нам хотелось бы увидеть в окне вывода (рис. 6.10).

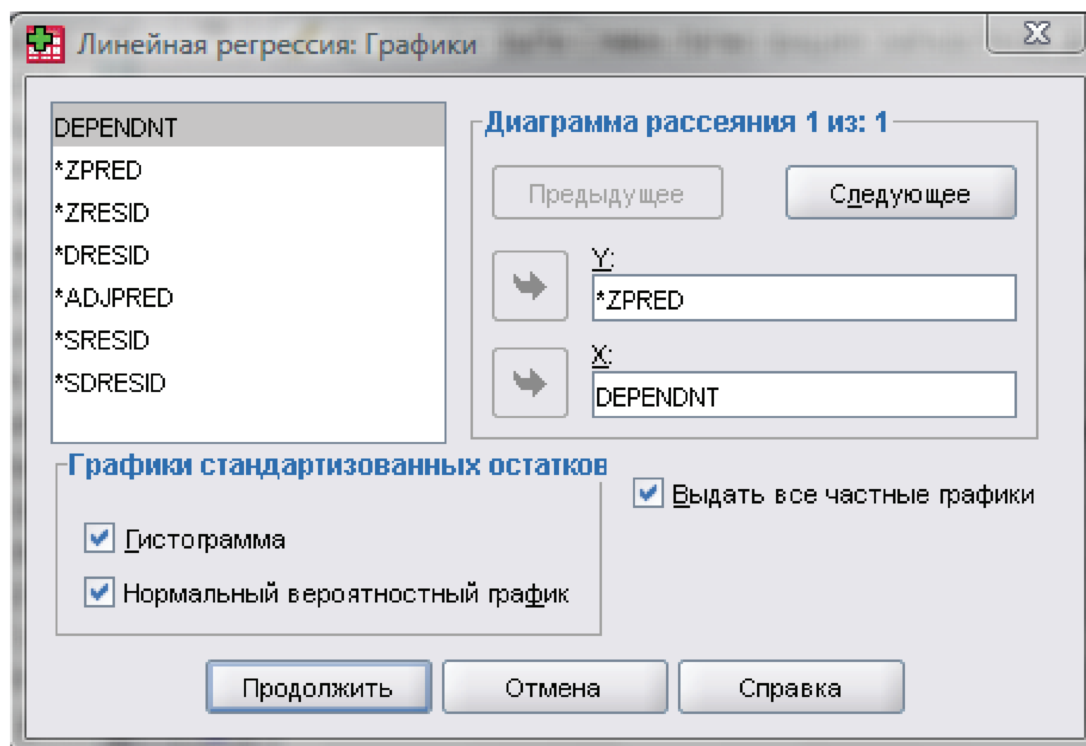


Рис. 6.10. Диалоговое окно «Линейная регрессия: графики»

Здесь:

- DEPENDNT зависимая переменная
- *ZPRED стандартизованные предсказанные значения
- *ZRESID стандартизованные остатки
- *DRESID удаленные остатки
- *ADJPRED скорректированные предсказанные значения
- *SRESID стьюдентизированные остатки
- *SDRESID стьюдентизированные удаленные остатки

После запуска команды построения графиков в окне вывода появляются следующие графики (рис. 6.11, 6.12, 6.13):



Рис. 6.11. Гистограмма



Рис. 6.12. Вероятностный график

Одним из простейших способов проверки нормальности распределения является построение гистограммы остатков, такой, как изображена

на рисунке 6.11. Здесь на гистограмму наблюденных частот (обозначенных столбиками) наложена кривая нормального распределения.

Другой способ сравнения эмпирического распределения остатков с распределением, ожидаемым при выполнении условия нормальности, состоит в выводе двух этих распределений (теоретическим и экспериментальным). Если распределения идентичны, они лягут на одну прямую линию. Наблюдая разброс точек вокруг прямой, соответствующей теоретическому нормальному закону, можно сравнить эти распределения.

На рисунке 6.13 представлена диаграмма рассеяния, которая представлена данными зависимой переменной по оси X и стандартизированными предсказанными значениями по оси Y (более подробно о стандартизированных остатках говорилось в лекции, посвященной таблицам сопряженности).



Рис. 6.13. Диаграмма рассеяния (надпись повтор + слева)

Виды регрессионного анализа, реализованные в SPSS

В этом разделе рассмотрим, какие виды регрессионного анализа существуют, и кратко охарактеризуем их.

Бинарная логистическая регрессия является полезной в ситуациях, когда необходимо предсказать результаты на основе значений из набора переменных. Она похожа на модель линейной регрессии, но подходит для моделей, где зависимой является дихотомическая переменная. Иными словами, с помощью метода бинарной логистической регрессии можно исследовать зависимость дихотомических переменных от независимых переменных, имеющих любой вид шкалы.

Исследователи часто хотят спрогнозировать произойдет или нет то или иное событие, например, голосование на выборах, участие в общественных программах, удача или неудача какого-либо бизнес-проекта и т.п. Бинарная логистическая регрессия может помочь в этом деле.

Приведем пример из области медицины, когда может быть использована логистическая регрессия. Какие характеристики образа жизни являются факторами риска ишемической болезни сердца (ИБС)? Зная такие параметры образа жизни пациентов, как: курение, диета, физические упражнения, употребление алкоголя, можно построить модель, способную предсказать наличие или отсутствие ИБС в выборке пациентов. Модель может быть использована для получения оценок шансов для каждого фактора. Так, например, можно предсказать, у каких пациентов более высокая вероятность заболеть ИБС – курящих или некурящих.

Мультиноминальная или полиноминальная логистическая регрессия может пригодиться в ситуациях, когда необходимо классифицировать предметы на основе значений из множества переменных предикторов. Этот тип регрессии похож на логистическую регрессию, но он является более общим, так как зависимая переменная не ограничивается двумя категориями.

Например, изучая рынок фильмов, киностудия может изучить целевую аудиторию, на которую нацелен тот или иной фильм. Выполняя полиномиальную логистическую регрессию, студия может определить взаимосвязь различных параметров потенциальных зрителей (возраст, пол, уровень дохода и др.) и типа фильма, который они предпочитают. В результате рекламная кампания фильма будет создаваться целенаправленно для той целевой аудитории, вероятность посещения фильма которой наибольшая.

Порядковая регрессия. В то время как мультиномиальная регрессия предназначена для зависимой переменной, относящейся к номинальной шкале, порядковая регрессия предназначена для целевой переменной, принадлежащей к порядковой шкале. Независимые переменные и здесь должны быть категориальными (то есть иметь номинальную или порядковую шкалу), однако в качестве ковариат допускается применение переменных с интервальной шкалой.

Например, при помощи порядковой регрессии можно изучить реакцию респондентов (участников фокус-групп, холл-теста и т.д.) на определенный товар. Возможные реакции можно классифицируем как «отсутствие», «слабая», «умеренная» или «сильная». Различие между «слабой» и «умеренной» реакциями оценить количественно сложно или невозможно – оно основано на восприятии. Более того, различие между «слабой» и «умеренной» реакциями может быть больше или меньше, чем различие между «умеренной»

и «сильной» реакциями. В этих случаях как раз и используется порядковая регрессия.

Пробит анализ – это вид регрессионного анализа, использующийся для определения влияния количественного признака на бинарный отклик.

Этот метод известен также под именем «Дозаторный анализ кривых воздействия» и находит применение преимущественно в области токсикологии. В большинстве случаев речь идет о том, как на заданное количество индивидуумов воздействуют различные дозировки некоторого вещества (к примеру, некоторого токсичного вещества).

Классический пример, который вошел и в справочник пакета SPSS, исследует действие средства, предназначенного для уничтожения насекомых. При этом производится подсчет, сколько насекомых из заранее известного количества погибли при воздействии определенных доз вещества. Особый интерес в данном случае представляет дозировка, при которой уничтожается половина имеющихся насекомых.

Можно привести и другой пример. Шеф секретной службы некоторой вымышленной страны пожелал узнать, сколько денег он должен предложить гражданам соседнего государства, чтобы они доставляли ему некоторую тайную информацию. Для этой цели через своего посредника он предлагает первой группе 1000 долларов и отмечает, сколько человек соглашаются на его предложение вести шпионскую деятельность. Второй группе он предлагает 2000 долларов, и вновь отмечает себе количество попаданий в цель. Он продолжает предлагать деньги и дальше, действуя таким пошаговым образом, и доходит до суммы 10000 долларов. При этом исследованиям подвергаются две различные категории людей. К первой категории относятся люди, которые недовольны своим материальным положением, ко второй – люди, удовлетворенные своим материальным положением.

Шеф секретной службы желает выяснить, сколько он должен предложить денег, чтобы достичь желаемой доли положительных ответов в обеих категориях. К примеру, его интересует сумма, которую он должен заплатить, чтобы на его предложение согласилась половина опрашиваемой группы.

Нелинейная регрессия. Все рассмотренные выше примеры – это линейные регрессионные модели, в которых переменные имели первую степень (модели, линейные по переменным), а параметры выступали в виде коэффициентов при этих переменных (модели, линейные по параметрам). Однако соотношение между социально-экономическими явлениями и процессами далеко не всегда можно выразить линейными функциями, так как при этом могут возникать неоправданно большие ошибки.

Так, например, нелинейными оказываются производственные функции (зависимости между объемом произведенной продукции и основными факторами производства — трудом, капиталом и т.п.), функции спроса (за-

висимость между спросом на товары или услуги и их ценами или доходом) и др.

Для оценки параметров нелинейных моделей используются два подхода. Первый основан на линеаризации модели и заключается в том, что с помощью подходящих преобразований исходных переменных исследуемую зависимость представляют в виде линейного соотношения между преобразованными переменными. Второй подход обычно применяется в случае, когда подобрать соответствующее линеаризующее преобразование не удастся. В этом случае применяются методы нелинейной оптимизации на основе исходных переменных.

Для линеаризации модели в рамках первого подхода могут использоваться как модели, не линейные по переменным, так и модели, не линейные по параметрам. Если модель нелинейна по переменным, то введением новых переменных ее можно свести к линейной модели, для оценки параметров которой мы используем обычный метод наименьших квадратов.

Взвешенная регрессия – это регрессия, оценки коэффициентов которой получают минимизацией взвешенной суммы квадратов остатков. Применяется для отражения новейших тенденций изучаемого явления. При сохранении формы регрессионного уравнения вводится неравноценное отношение к ошибкам (остаткам) уравнения в начале и в конце выборочного периода: старые ошибки получают меньший вес, а ошибкам последних моментов придаются большие веса. Таким образом, веса становятся функцией времени. Эта функция характеризует «память» модели. Если изучаемый процесс претерпевает быстрые изменения, то и весовая функция должна быстро убывать (затухать) при движении от текущего момента в прошлое. Оптимальность весовой функции определяется минимумом суммы квадратов ошибок в прогнозах на ретроспективных данных.

Приведем такой пример. Допустим, прогнозируется вес ребенка в зависимости от его возраста. Ясно, что дисперсия веса для четырехлетнего младенца будет значительно меньше, чем дисперсия веса 14-летнего юноши. Проблема неоднородности дисперсии в регрессионном анализе называется проблемой гетероскедастичности.

В SPSS имеется возможность корректно сделать соответствующие оценки за счет приписывания весов слагаемым минимизируемой суммы квадратов. Естественно, чем меньше дисперсия остатка на объекте, тем больший вес он будет иметь. В приведенном примере на достаточно больших данных можно оценить дисперсию для каждой возрастной группы и вычислить необходимую весовую переменную. Увеличение влияния возрастных групп с меньшим возрастом в данном случае вполне оправдано.

Двухэтапный метод наименьших квадратов. Стандартные модели линейной регрессии предполагают, что ошибки в зависимой переменной

не связаны с независимой переменной или переменными, в случае если их несколько. Если это не так (например, когда отношения между переменными являются двунаправленными), используемый линейной регрессией метод наименьших квадратов (МНК) уже не обеспечивает оптимальные оценки модели. Двухэтапный метод наименьших квадратов регрессии использует инструментальные переменные, которые коррелируют с ошибками, чтобы вычислить расчетные значения проблематичного предиктора (первый этап), а затем уже использует эти расчетные значения для оценки модели линейной регрессии зависимой переменной (второй этап). Поскольку расчетные значения основаны на переменных, которые коррелируют с ошибками, результаты модели двухступенчатого являются оптимальными.

Допустим, необходимо выяснить, будет ли зависеть спрос на определенный товар от его цены и дохода потребителя. Проблемным моментом в этой модели является то, что цена и спрос оказывают обратное влияние друг на друга. То есть цены могут влиять на спрос и спрос может также влиять на цену. Регрессия, основанная на двухэтапном методе наименьших квадратов, может использовать доход потребителей и цену для расчета такого значения цены, которое не будет зависеть от изменений спроса. Этот новое значение цены в процессе расчета заменяет саму цену в первоначально указанной модели.

Категориальная регрессия предсказывает значения зависимой категориальной переменной по комбинации независимых категориальных переменных.

Данная процедура осуществляет оцифровку категориальных данных, путем присвоения категориям числовых значений, и построение оптимального уравнения регрессии для преобразованных данных. Категориальная регрессия может применяться, например, для описания удовлетворенности покупателей в зависимости от простоты совершения покупки, цены и качества товара. Полученное уравнение можно использовать с целью предсказания удовлетворенности покупателей для любых сочетаний значений независимых переменных.

Другим примером может послужить ситуация, когда мы хотим выяснить, каким образом удовлетворенность выполняемой работой зависит от вида деятельности, региона и количества командировок. Может оказаться, что высокие уровни удовлетворенности характерны для менеджеров и сотрудников, редко отправляемых в командировки. Результирующее уравнение регрессии может использоваться для предсказания степени удовлетворенности работой по комбинациям трех независимых переменных.

Вопросы и задания

1. Дайте общую характеристику регрессионного анализа данных и его возможностей.
2. Какие статистические показатели используются в SPSS при работе с линейной регрессией?
3. Опишите формы графического изображения и принципы построения графиков и диаграмм в регрессионном анализе.
4. Перечислите и охарактеризуйте виды регрессионного анализа, реализованного в SPSS.
5. Самостоятельно осуществите процедуру линейной регрессии, используя данные ЕАЭСД.

Список литературы

1. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цёфель. – СПб.: ДиаСофтЮП, 2005. – 608 с.

2. Крыштановский, А.О. Анализ социологических данных / А.О. Крыштановский – М.: Изд-во «ГУ ВШЭ», 2007. – 281 с.

3. Езекиел, М. Методы анализа корреляций и регрессий / М. Езекиел, К. Фокс. – М.: Статистика, 1966. – 361 с.

4. Турундаевский, В. Б. Многомерный статистический анализ в экономических задачах. Компьютерное моделирование в SPSS / В. Б. Турундаевский, И. В. Орлова, Н. А. Концевая. – М.: Изд-во «Вузовский учебник», 2009. – 320 с.

Лекция 7. Сравнение средних

Метод сравнения средних наиболее часто применяется при статистическом анализе данных. Это связанно с тем, что анализ данных начинается с группировки и вычисления описательных статистик в группах, например, вычисления средних и стандартных отклонений. Если в распоряжении исследователя имеется две группы данных, то естественно сравнить средние значения в этих группах. Такого рода задачи во множестве возникают на практике, например, мы можем сравнить средний доход различных групп людей: проживающих в столице, обычном городе или сельской местности, для того чтобы оценить дифференциацию уровня жизни.

В 17-й версии программы SPSS существует несколько видов анализа объединенных в общую группу «Сравнение средних» (рис. 7.1). Остановимся на каждом из них более подробно

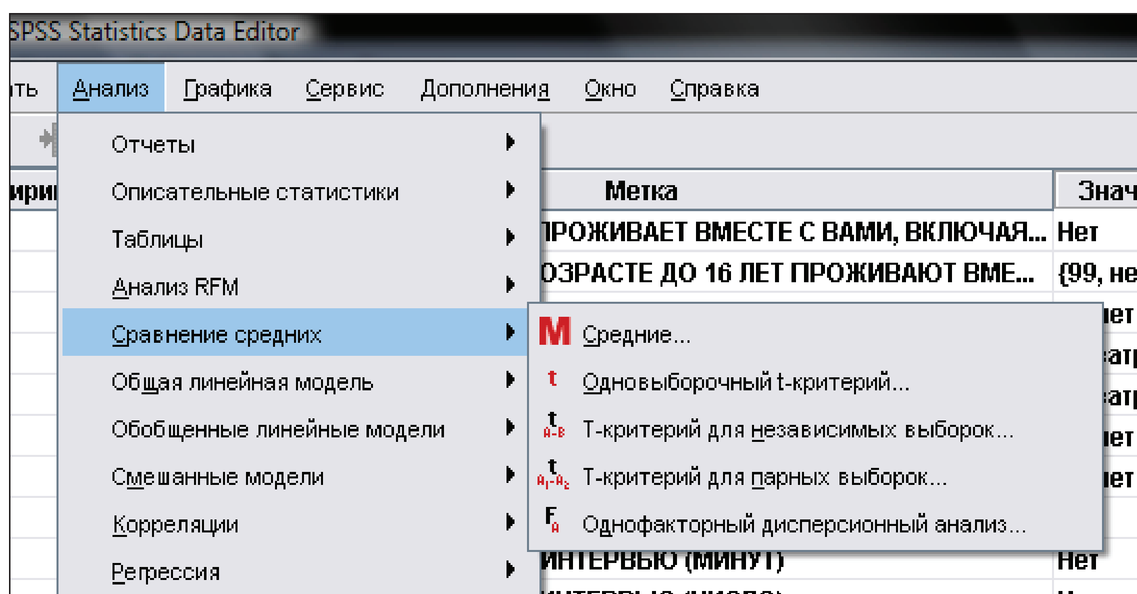


Рис. 7.1. Сравнение средних

М средние

Процедура «М средние» вычисляет средние значения для подгрупп и связанные с ними одномерные статистики для зависимых переменных внутри категорий одной или нескольких независимых переменных.

Данные для процедуры «М средние» должны быть следующие: зависимые переменные – количественные, независимые переменные – категориальные.

Процедура «М средние» осуществляется вызовом диалогового окна следующего вида (рис.7.2):

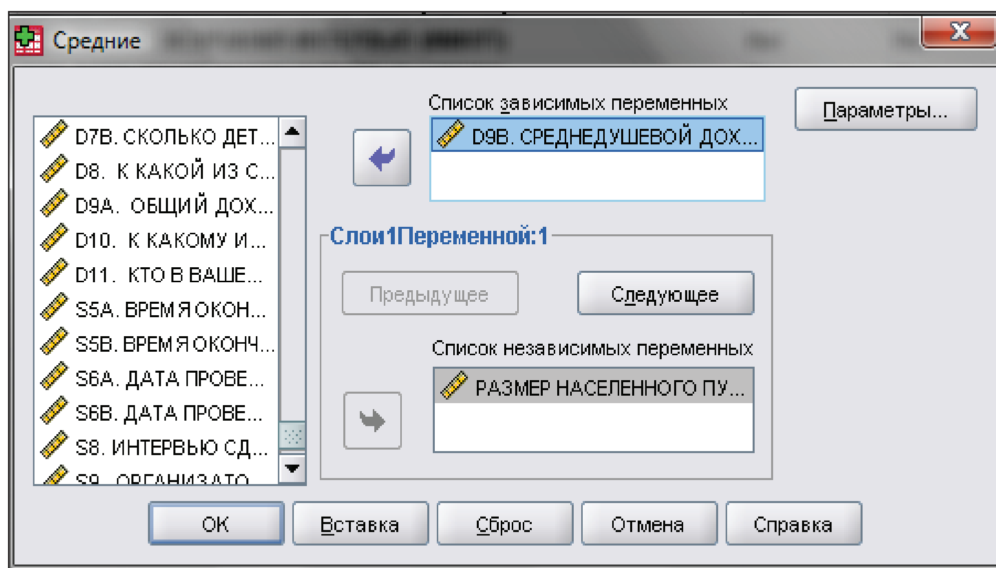


Рис. 7.2. Диалоговое окно «Средние»

В качестве примера сравним среднедушевой доход жителей различных по численности населенных пунктов. Вводим в окно «Список зависимых переменных» переменную «Среднедушевой доход», а в окно «Список независимых переменных» – «Размер населенного пункта по пяти позициям»[18].

Далее нажимаем кнопку «Параметры» для выбора одной или нескольких видов статистик для подгрупп, рассчитываемых для переменных внутри каждой отдельной категории каждой группирующей переменной. Таковыми, как видно из рис. 7.3, являются: сумма, число наблюдений, среднее значение, медиана, медиана группы, стандартная ошибка среднего значения, минимальное и максимальное значения, размах, значение группирующей переменной для первой категории, значение группирующей переменной для последней категории, стандартное отклонение, дисперсия, эксцесс и др. Здесь же можно изменить порядок, в котором выводятся статистики подгрупп. Порядок, в котором статистики приведены в списке «Статистики» в ячейках, определяет их порядок при выводе. Итожащие статистики также выводятся для каждой переменной по всем категориям.

Кроме того, в нижней части этого же диалогового окна имеется возможность выбрать «Статистики для первого слоя», которые включают в себя следующие показатели:

- Таблица дисперсионного анализа и Эта. Выводит таблицу однофакторного дисперсионного анализа и вычисляет значение Эта и Эта в квадрате (меры близости) для каждой независимой переменной в первом слое.
- Критерий линейности. Вычисляет сумму квадратов, степени свободы и средний квадрат для линейных и нелинейных компонентов, а также F-отношение, значения R и R-квадрат.

[18] Использовалась база данных «Курьер», 2010 г., 12-ая волна.

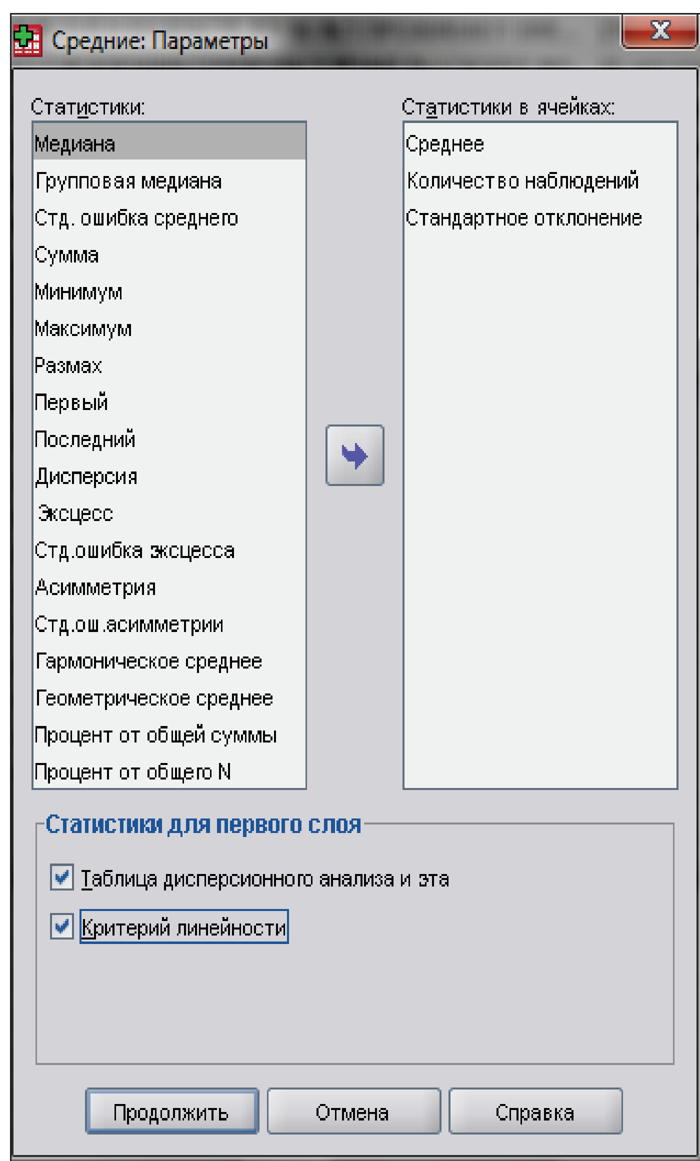


Рис. 7.3. Диалоговое окно «Средние: Параметры»

Здесь необходимо отметить следующее. Некоторые статистики для подгрупп, например, среднее и стандартное отклонение, основаны на теории нормального распределения и подходят для количественных переменных с симметричными распределениями. Робастные[19] статистики, такие

[19] Под робастностью в статистике понимают нечувствительность к различным отклонениям и неоднородностям в выборке, связанным с теми или иными, в общем случае неизвестными, причинами. Это могут быть ошибки детектора, регистрирующего наблюдения, чьи-то добросовестные, или не добросовестные, попытки «подогнать» выборку до того, как она попадёт к статистику, ошибки оформления, вкравшиеся опечатки и многое другое. Робастность в статистике предоставляет подходы, направленные на снижение влияния выбросов и других отклонений в исследуемой величине от моделей, используемых в классических методах статистики. Источник: Википедия – свободная интернет-энциклопедия, http://ru.wikipedia.org/.

как медиана, годятся и для количественных переменных, которые могут не удовлетворять условию нормального распределения. Дисперсионный анализ является робастным в отношении отклонений от нормальности, однако данные в каждой ячейке должны быть симметричными. При проведении дисперсионного анализа предполагается, что группы принадлежат совокупностям с одинаковыми дисперсиями.

Для примера выберем статистические показатели «Средние», «Количество наблюдений», «Стандартное отклонение». А также выделим показатели: «Таблица дисперсионного анализа и Эта» и «Критерий линейности». В результате программа выводит следующие данные:

Таблица 7.1

Сводка обработки наблюдений

	Наблюдения					
	Включенные		Исключенные		Итого	
	N	Процент	N	Процент	N	Процент
D9B. СРЕДНЕДУШЕВОЙ ДОХОД * РАЗМЕР НАСЕЛЕННОГО ПУНКТА ПО ПЯТИ ПОЗИЦИЯМ	1191	74,4%	409	25,6%	1600	100,0%

Первая таблица показывает общие данные по наблюдениям. Из данных таблицы видно, что для процедуры «М средние» было отобрано 1191 наблюдений, что составляет 74,4% от общего числа респондентов. Соответственно 25, 6% было исключено – это респонденты, которые отказались указывать свой доход.

В следующей таблице представлен отчет по средним значениям для каждой подвыборки (табл. 7.2)

Таблица 7.2

Отчет

РАЗМЕР НАСЕЛЕННОГО ПУНКТА ПО ПЯТИ ПОЗИЦИЯМ	Среднее	N	Стд.отклонение
Москва	16164,26	70	7146,615
более 500 тыс.	9574,87	240	4725,278
от 100 до 500 тыс.	8188,30	227	4447,189
города до 100 тыс.	7361,25	324	4552,564
село	6415,59	330	3829,544
Итого	8220,32	1191	5105,739

Во втором столбце («Среднее») видны средние значения среднедушевого дохода. В третьем столбце («N») представлена численность респонден-

тов, на основе которых вычислялось среднее значение. Последний столбец («Стандартное отклонение») показывает среднее отклонение от среднего значения выборки.

Третья таблица – таблица дисперсионного анализа ANOVA (английская аббревиатура ANOVA происходит от Analysis of VariaTion – Дисперсионный анализ). Дисперсионный анализ, который рассматривает только одну переменную, называется однофакторным дисперсионным анализом (One-Way ANOVA). Дисперсионный анализ может также применяться в случае двух переменных – это двухфакторный дисперсионный анализ (Two-Way ANOVA). В нашем примере речь будет идти об однофакторном анализе.

Таблица 7.3

Таблица ANOVA

			Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
D9B. СРЕДНЕДУШЕВОЙ ДОХОД * РАЗМЕР НАСЕЛЕННОГО ПУНКТА ПО 5-ТИ ПОЗИЦИЯМ	Между группами	(Комбинированная)	6,172E9	4	1,543E9	73,642	,000
		Линейность	4,548E9	1	4,548E9	217,067	,000
		Отклонение от линейности	1,624E9	3	5,413E8	25,834	,000
	В группах		2,485E10	1186	2,095E7		
	Итого		3,102E10	1190			

Самыми важными показателями в этой таблице являются уровень значимости (последний столбец «Знч.») и критерий, отношения среднего квадрата между группами к среднему квадрату внутри группы (предпоследний столбец «F»). Эти два показателя необходимо рассматривать вместе. Когда значение F велико и уровень значимости мал (как правило меньше, чем 0,05 или 0,01) нулевая гипотеза может быть отвергнута. Другими словами, небольшой уровень значимости означает, что результаты, вероятно, не случайны. В нашем случае p = 0,000. Это указывает на то, что разность между средними значениями переменной для всех групп статистически достоверна. Ниже дана трактовка остальных показателей, используемых программой в окне вывода.

Значение в столбце «Сумма квадратов» строки «Между группами» означает сумму квадратов разностей между общим средним значением и средними значениями каждой группы, умноженными на весовые коэффициенты, равные числу объектов в группе, а строка «В группах» сумму квадратов разностей среднего значения каждой группы и каждого значения этой группы.

Значение в столбце «Ст. св.» строки «Между группами» означает межгрупповое число степеней свободы, равное числу групп, уменьшенному на единицу, в строке «В группах» – внутригрупповое число степеней свободы, равное разности между числом объектов и числом групп.

«Средний квадрат» – отношение суммы квадратов к числу степеней свободы.

Последняя таблица в окне вывода показывает меры связи между переменными (табл.7.4).

Таблица 7.4

Меры связи

	R	R квадрат	Эта	Эта квадрат
D9B. СРЕДНЕДУШЕВОЙ ДОХОД * РАЗМЕР НАСЕЛЕННОГО ПУНКТА ПО ПЯТИ ПОЗИЦИЯМ	,383	,147	,446	,199

R – коэффициент корреляции между наблюдаемыми и предсказанными значениями зависимой переменной. Она варьируется в диапазоне от 0 до 1. Малая величина показывает, имеет место малая линейная зависимость между переменными, или ее нетсовсем.

R-квадрат – мера согласия или мера линейной модели, которую иногда называют коэффициентом детерминации. Это доля изменения зависимой переменной регрессионной модели. Она (кто?) варьируется в диапазоне от 0 до 1. Малые значения показывают, что модель плохо вписывается в данные.

Эта – мера ассоциации, которая колеблется в диапазоне от 0 до 1. При этом 0 означает отсутствие связи между переменными а значение, близкое к 1, указывает на высокую степень ассоциации.

Эта-квадрат – мера ассоциации, которая подходит для зависимой переменной, измеряемой интервальной шкалой, и независимой переменной с ограниченным числом категорий. Эта является асимметричным показателем и не связана с линейной зависимостью между переменными. Показатель Эта-квадрат можно интерпретировать как долю дисперсии в зависимой переменной, объясняющей различия между группами.

Исходя из всего вышесказанного, осуществим смысловую интерпретацию полученных результатов. Итак, вторая таблица показывает нам уменьшение средних значений среднедушевого дохода, и, соответственно, его колебаний по мере уменьшения численности населенного пункта. Следующая таблица говорит нам о статистической достоверности полученных результатов, а последняя таблица показывает, что связь между переменными, хоть и не очень сильная, но имеется. То есть мы можем смело делать вывод: чем выше численность в населенном пункте, тем выше уровень среднедушевого дохода семей его жителей.

Одновыборочный Т-критерий

Процедура Одновыборочный Т-критерий проверяет, отличается ли среднее одной переменной от заданной константы.

Допустим, требуется узнать, отличается ли средний IQ группы студентов от 100. Или, например, при демографическом исследовании можно выявить, отличается ли среднее количество членов семьи россиян от 3 человек при 95% доверительном уровне.

Для каждой проверяемой переменной можно выявить следующие показатели: среднее значение, стандартное отклонение и стандартную ошибку среднего значения, среднюю разность между каждым значением данных и гипотетической проверяемой величиной, Т-критерий для проверки равенства этой разности нулю, доверительный интервал для этой разности (доверительный уровень можно задать самому).

Чтобы выполнить тест для значений количественной переменной и гипотетического проверяемого значения, необходимо выбрать количественную переменную и ввести гипотетическое проверяемое значение.

Т-критерий предполагает, что данные нормально распределены; однако, он довольно устойчив к отклонениям от нормальности.

Итак, осуществим эту процедуру на конкретном примере. Вызываем диалоговое окно «Одновыборочный Т-критерий» (рис.4). Вводим в окно «Проверять переменные» изучаемую переменную, в нашем случае это будет «D7A. Сколько человек проживает вместе с Вами, включая Вас лично и всех детей?». Вводим «Проверяемое значение», равное 3. Таким образом, мы попытаемся выяснить, на сколько среднее количество членов семьи россиян отличается от значения 3.

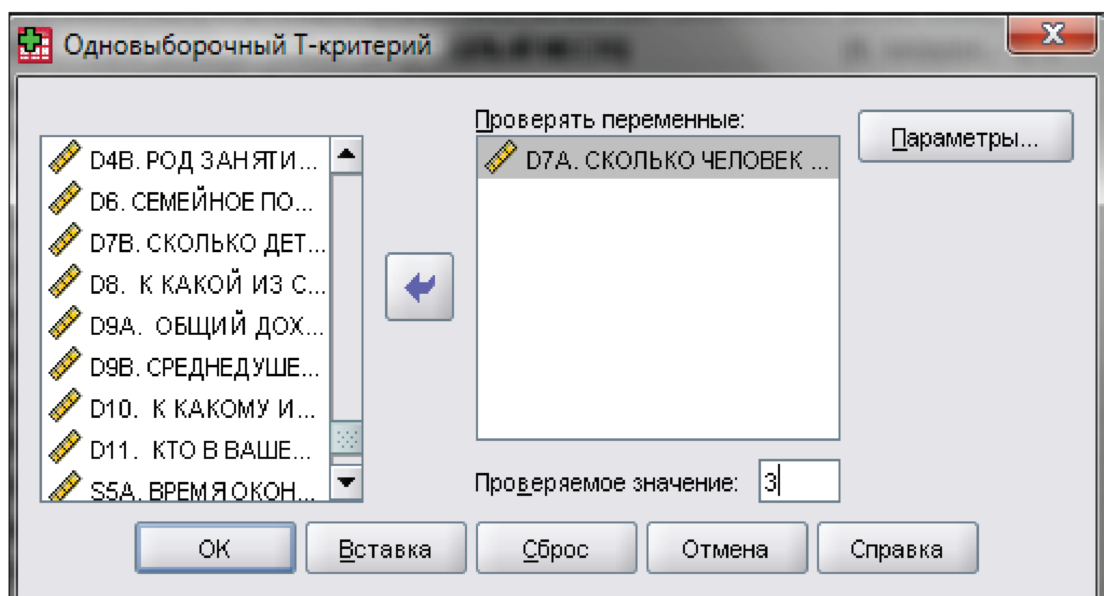


Рис. 7.4. Диалоговое окно «Одновыборочный Т-критерий»

Нажав на кнопку «Параметры» вызываем дополнительное диалоговое окно (рис. 7.5).

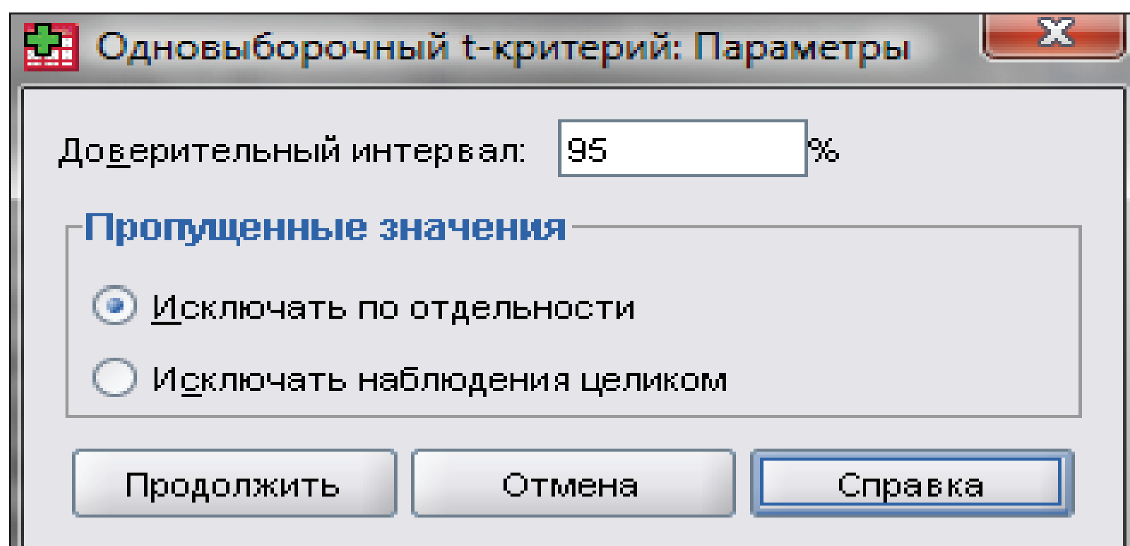


Рис. 7.5. Диалоговое окно «Одновыборочный Т-критерий: Параметры»

Здесь мы можем задать необходимый нам доверительный интервал[20]. По умолчанию для разности среднего и гипотетического проверяемого значения выводится 95%-й доверительный интервал. Чтобы задать другой доверительный уровень, необходимо ввести значение между 1 и 99.

Когда проверяется несколько переменных, и некоторые из них содержат пропущенные значения, в разделе «Пропущенные значения» можно указать, какие наблюдения следует включить (или исключить):

1. **Исключать по отдельности.** При применении Т-критерия используются все наблюдения, в которых проверяемые переменные имеют непропущенные значения. Объемы выборок могут меняться в зависимости от переменных, к которым применяется критерий.
2. **Исключать наблюдения целиком.** Каждый раз при применении Т-критерия используются только те наблюдения, которые не имеют пропущенных значений для всех переменных, для которых запрошено применение Т-критерия. Объем выборок одинаков для всех тестов.

В этом диалоговом окне оставим значения, заданные программой по умолчанию. В результате в окне вывода программа представляет нам следующие результаты:

---

[20] Доверительный интервал – это допустимое отклонение наблюдаемых значений от истинных. Размер этого допущения определяется исследователем с учетом требований к точности информации.

Таблица 7.5

Статистики для одновыборочного Т-критерия

	N	Среднее	Стд. отклонение	Стд. ошибка среднего
D7A. СКОЛЬКО ЧЕЛОВЕК ПРОЖИВАЕТ ВМЕСТЕ С ВАМИ, ВКЛЮЧАЯ ВАС ЛИЧНО И ВСЕХ ДЕТЕЙ?	1600	2,90	1,307	,033

В табл. 7.5 представлены общие статистики для одновыборочного Т-критерия: число наблюдений N=1600 респондентов, среднее значение 2,90, стандартное отклонение 1,307 (показывает степень разброса отдельных индивидуальных наблюдений относительно этого среднего) и стандартная ошибка среднего 0,33 (теоретическое стандартное отклонение всех средних выборки).

Вторая таблица показывает непосредственно результаты процедуры «Одновыборочный Т-критерий» (табл.7.6).

Таблица 7.6

Одновыборочный Т-критерий

	Проверяемое значение = 3					
	Т	ст.св.	Значимость (двухсторонняя)	Разность средних	95%-й доверительный интервал разности средних	
					Нижняя граница	Верхняя граница
D7A. СКОЛЬКО ЧЕЛОВЕК ПРОЖИВАЕТ ВМЕСТЕ С ВАМИ, ВКЛЮЧАЯ ВАС ЛИЧНО И ВСЕХ ДЕТЕЙ?	–3,022	1599	,003	–,099	–,16	–,03

Здесь во втором столбце таблицы отображается статистика Т для каждого наблюдения, которая рассчитывается как отношение разницы средних к стандартной ошибке выборочного среднего. Отличие критерия от 0 выражает различие среднего значения переменной от эталонной величины.

В следующем столбце отображаются число степеней свободы, равное разности между числом объектов в группе минус единица.

«Значимость двухсторонняя» показывает вероятность от Т распределения с имеющимися степенями свободы. Статистическая значимость не больше 0,05 считается значимой.

Следующий столбец, «Разность средних», получается путем вычитания тестового значения от выборочного среднего.

Последний столбец указывает на значения верхней и нижней границы, входящие в 95%-й доверительный интервал.

Интерпретируя полученные данные, отметим следующее: среднее значение количества человек в семье россиян близко к проверяемому значению – 3, однако велико и отклонение от средней. Данный вывод можно подтвердить рассмотрев частотное распределение изучаемой переменной (табл. 7.7).

Таблица 7.7

Частотное распределение переменной «D7A. СКОЛЬКО ЧЕЛОВЕК ПРОЖИВАЕТ ВМЕСТЕ С ВАМИ, ВКЛЮЧАЯ ВАС ЛИЧНО И ВСЕХ ДЕТЕЙ?»

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	1	213	13,3	13,3	13,3
	2	455	28,4	28,4	41,8
	3	451	28,2	28,2	69,9
	4	311	19,4	19,4	89,4
	5	124	7,8	7,8	97,1
	6	31	1,9	1,9	99,1
	7	10	,6	,6	99,7
	8	1	,1	,1	99,8
	9	2	,1	,1	99,9
	10	2	,1	,1	100,0
	Итого	1600	100,0	100,0	

Т-критерий для независимых выборок

Процедура Т-критерий для независимых выборок сравнивает средние значения для двух групп наблюдений. В идеале объекты для этого критерия должны быть случайным образом приписаны двум группам, чтобы любое различие в отклике определялось рассматриваемым воздействием, например, лечением (или его отсутствием), а не другими факторами. Это не выполняется, если вы сравниваете средний доход для мужчин и женщин. Пол не приписывается индивидууму случайным образом. В подобных ситуациях следует убедиться, что различия в других факторах не снижают и не увеличивают значимые различия средних значений. Например, на различие средних доходов кроме пола может оказывать влияние такой фактор, как образование.

Типичный пример для иллюстрации процедуры Т-критерий для независимых выборок, который приведен в разделе справки пакета SPSS, выглядит следующим образом:

Пациенты с высоким давлением случайным образом делятся на контрольную группу и группу испытуемых. Пациенты в контрольной группе получают плацебо (фармакологически неактивные таблетки), а пациенты в группе испытуемых получают лекарство (исследуемые таблетки, которые предположительно понижают давление). Пациенты наблюдаются в течение двух месяцев, после чего для сравнения средних значений кровяного давления пациентов контрольной группы и группы испытуемых применяют двухвыборочный Т-критерий. Давление каждого пациента измеряют один раз, и каждый пациент принадлежит только к одной группе.

В процедуре Т-критерий для независимых выборок проверяется количественная переменная. Чтобы разбить наблюдения на две группы, используется группирующая переменная с двумя значениями. Эта переменная может быть числовой (например, со значениями 1 и 2 или 6,25 и 12,5) или короткой текстовой (например, со значениями «да» и «нет»). В случае, если группирующая переменная содержит несколько значений (например, «образование»), из всех ее значений можно выбрать два, по которым и будет производиться группировка. Можно также использовать количественную переменную, такую как возраст, чтобы разбить наблюдения на две группы путем задания пороговой точки (пороговая точка «21» разбивает возраст на группы: до 21 года и 21 год или более).

Рассмотрим изучаемую процедуру на конкретном примере. В базе данных «Курьер», 2010 г., 10-я волна имеются следующие переменные – количественная «D9B. Среднедушевой доход» и номинальная «D10. К какому из следующих социальных слоев Вы бы отнесли себя и свою семью?» с со следующими значениями:

0 = «нет ответа»
1 = «высший слой»
2 = «верхняя часть среднего слоя»
3 = «средняя часть среднего слоя»
4 = «нижняя часть среднего слоя»
5 = «низший слой»

Попытаемся осуществить процедуру Т-критерий для независимых выборок, сравнив между собой средние значения дохода для различных представителей среднего слоя, в частности, нижней и верхней ее частей. Для этого вызываем диалоговое окно «Т-критерий для независимых выборок» (рис 7.6)

В появившемся диалоговом окне вводим в окно «Проверяемые переменные» количественную переменную «D9B. Среднедушевой доход», а в диалоговое окно «Группировать по:» номинальную переменную «D10. К какому из следующих социальных слоев Вы бы отнесли себя и свою семью?» (рис. 7.7).

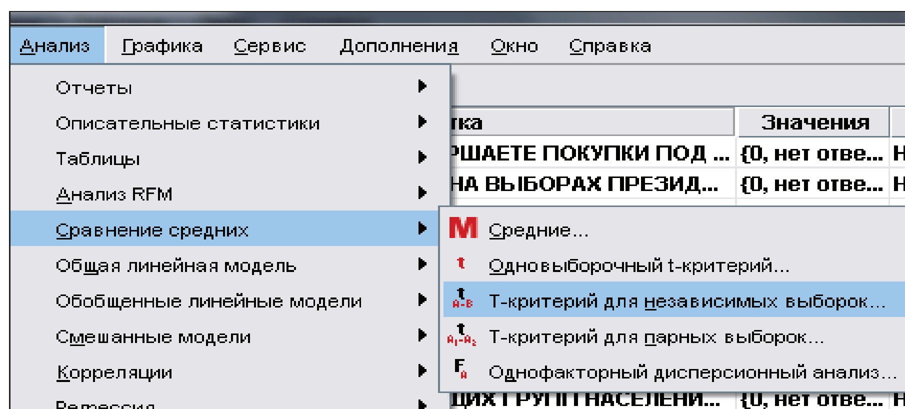


Рис. 7.6. Вызов диалогового окна «Т-критерий для независимых выборок»

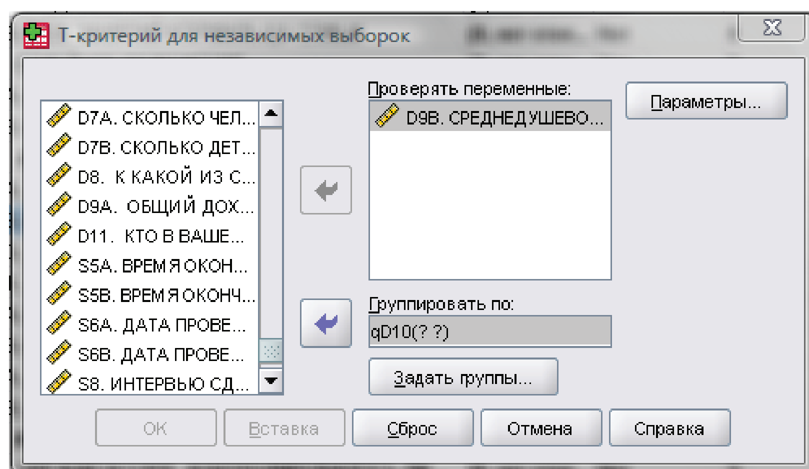


Рис. 7.7. Диалоговое окно «Т-критерий для независимых выборок»

Далее нажимаем на кнопку «Задать группы…», в результате чего появляется еще одно окно, в котором вводим значения сравниваемых выборок. В нашем случае это будет:

2 = «верхняя часть среднего слоя»

4 = «нижняя часть среднего слоя»

После этого нажимаем кнопку «Продолжить» (рис. 7.8).

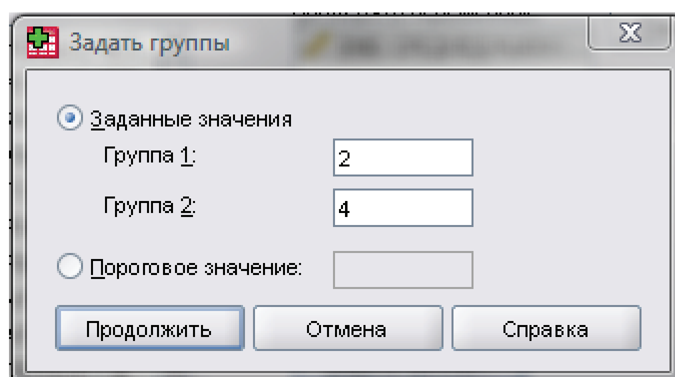


Рис. 7.8. Диалоговое окно «Задать группы»

Для того чтобы задать параметры процедуры, вызываем диалоговое окно «Т-критерий для независимых выборок: параметры».

Параметры процедуры Т-критерий для независимых выборок включают в себя «доверительный интервал». По умолчанию для разности средних значений выводится 95%-й доверительный интервал. Чтобы задать другой доверительный уровень, необходимо ввести значение между 1 и 99.

В разделе «Пропущенные значения» в случае, если проверяется несколько переменных и некоторые из них содержат пропущенные значения, можно указать, какие наблюдения следует включить (или исключить).

- «Исключать по отдельности». При работе с Т-критерием используются все наблюдения, в которых проверяемая переменная имеет непропущенные значения. Объемы выборок могут меняться в зависимости от переменных, к которым применяется критерий.
- «Исключать наблюдения целиком». При каждом применении Т-критерия используются только те наблюдения, которые не имеют пропущенных значений для всех переменных, для которых запрошено применение Т-критерия. Объем выборок одинаков для всех тестов (рис.7.9).

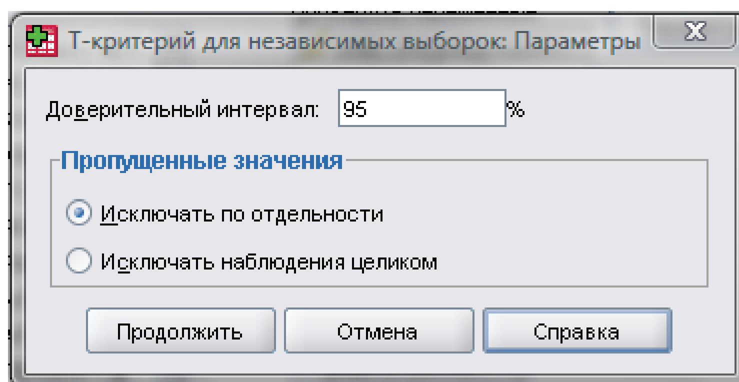


Рис. 7.9. Задача параметров в диалоговом окне «Т-критерий для независимых выборок: параметры»

Оставляем в этом окне параметры, установленные программой по умолчанию, то есть доверительный интервал 95% и исключение пропущенных значений по отдельности, нажимаем «Продолжить».

В итоге, при нажатии кнопки «ОК», в диалоговом окне «Т-критерий для независимых выборок», окно вывода выводит для пользователя две таблицы (табл. 7.8 и 7.9).

Практически все показатели, представленные в таблицах, нам уже знакомы – о них было сказано выше. Однако, здесь появился еще один критерий, с которым нам еще не приходилось сталкиваться, – «Критерий равенства дисперсий Ливиня». Он рассчитывается программой SPSS автоматически при расчете Т-критерия и является более устойчивым к нарушению нор-

мальности распределения. По каждому случаю он вычисляет абсолютную разницу между значением этого случая и его средним значением, а также выполняет однофакторный дисперсионный анализ по этим различиям.

Таблица 7.8

Групповые статистики

	D10. К КАКОМУ ИЗ СЛЕДУЮЩИХ СОЦИАЛЬНЫХ СЛОЕВ ВЫ БЫ ОТНЕСЛИ СЕБЯ И СВОЮ СЕМЬЮ?	N	Среднее	Стд. отклонение	Стд. ошибка среднего
D9B. СРЕДНЕДУШЕВОЙ ДОХОД	верхняя часть среднего слоя	41	7391,83	7130,729	1113,633
	нижняя часть среднего слоя	605	5964,86	5101,686	207,413

Таблица 7.9

Критерий для независимых выборок

		Критерий равенства дисперсий Ливиня		Т-критерий равенства средних						
									95%-й доверительный интервал разности средних	
		F	Знч.	Т	ст.св.	Значимость (двухсторонняя)	Разность средних	Стд. ошибка разности	Нижняя граница	Верхняя граница
D9B. СРЕДНЕДУШЕВОЙ ДОХОД	Предполагается равенство дисперсий	16,020	,000	1,684	644	,093	1426,966	847,336	-236,909	3090,842
	Равенство дисперсий не предполагается			1,260	42,820	,215	1426,966	1132,783	-857,787	3711,720

Нулевая гипотеза, которую проверяет критерий Ливиня, – равенство внутригрупповых дисперсий (в соответствии с нулевой гипотезой, дисперсии двух совокупностей равны). Показателями этого критерия являются:

F – значение критерия,

Знч. – уровень значимости.

Если уровень значимости меньше 0,05, что мы наблюдаем в нашем случае, то нулевая гипотеза о равенстве дисперсий должна быть отвергнута. В этом случае для интерпретации результатов можно использовать только вторую строку таблицы.

Интерпретируем полученный результат. Средние значения среднедушевого дохода представителей верхней и нижней части среднего слоя отличаются между собой (7130,729 и 5101,686 соответственно). Разброс уровня

дохода относительно среднего отличается в рассматриваемых группах населения, и у представителей верхней части среднего слоя он существенно выше. Полученные данные можно объяснить тем, что в современном российском обществе отсутствует четкая самоидентификация населения с принадлежностью к тому или иному стратификационному слою, причем у тех, кто относит себя к среднему высшему слою, она проявляется в большей степени.

Т-критерий для парных выборок

Процедура Т-критерий для парных выборок сравнивает средние значения переменных для одной группы наблюдений. Для всех наблюдений вычисляются разности значений двух переменных, а затем проверяется, отличается ли среднее этих разностей от нуля.

Типичный пример для иллюстрации процедуры Т-критерий для независимых выборок, который приведен в разделе справки пакета SPSS, выглядит следующим образом. При изучении проблемы повышенного артериального давления измеряют артериальное давление всем пациентам, проводят лечение, а затем повторно измеряют давление. Таким образом, для каждого пациента измерения проводят два раза (такие измерения часто называют измерениями «до» и «после»). Альтернативным планом эксперимента для применения этого критерия является исследование пар сочетаемых индивидуумов, или исследование типа «случай-контроль». При изучении кровяного давления пациенты и соответствующие контрольные субъекты могут подбираться по возрасту (75-летнему пациенту соответствует 75-летний член контрольной группы).

В социологических исследованиях процедура Т-критерий для парных выборок может использоваться для анализа количественных данных, полученных в результате мониторинговых исследований, проводимых на панельных выборках. Например, можно сравнить рейтинги политических деятелей в период предвыборной кампании. К сожалению, подобные данные достаточно тяжело найти в архивах социологических исследований, поэтому не будем приводить подробный пример реализации процедуры Т-критерия для парных выборок. Тем более что все статистики, используемые при ее реализации, – те же, что и для остальных процедур сравнения средних. Рассмотрим лишь алгоритм осуществления процедуры в SPSS.

Диалоговое окно вызывается уже известным нам способом (рис. 7.10).

В окне из общего списка переменных переносим в окно «Парные переменные» те, которые нам необходимо проанализировать (рис.7.11).

Вызвав диалоговое окно «Параметры», можно обнаружить все те же параметры, что использовались и для независимых парных выборок (рис. 7.12).

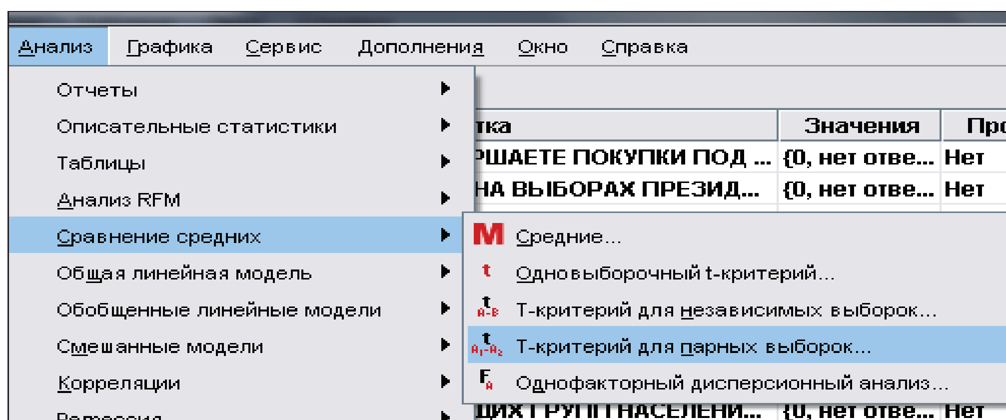


Рис. 7.10. Вызов диалогового окна «Т-критерий для парных выборок»

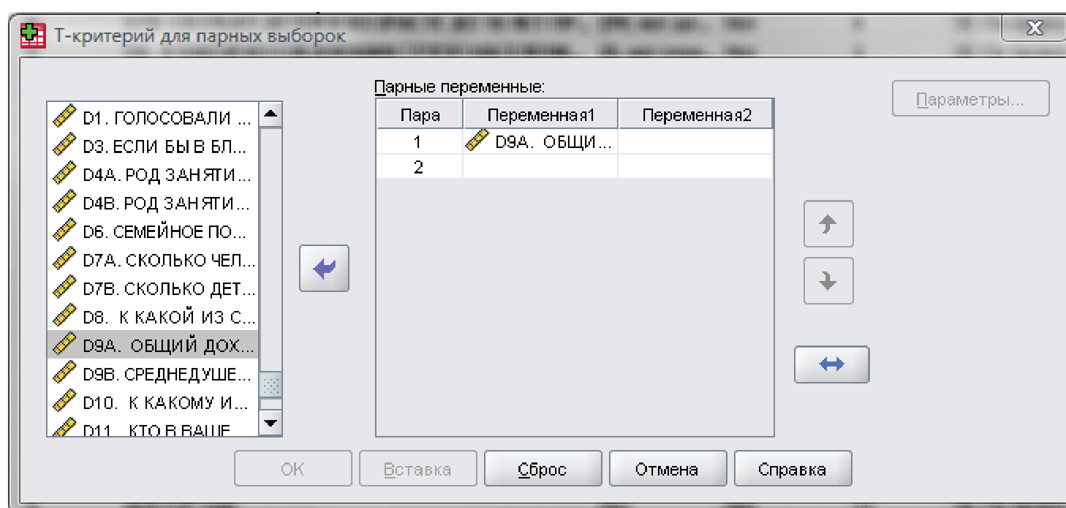


Рис. 7.11. Диалогового окно «Т-критерий для парных выборок»

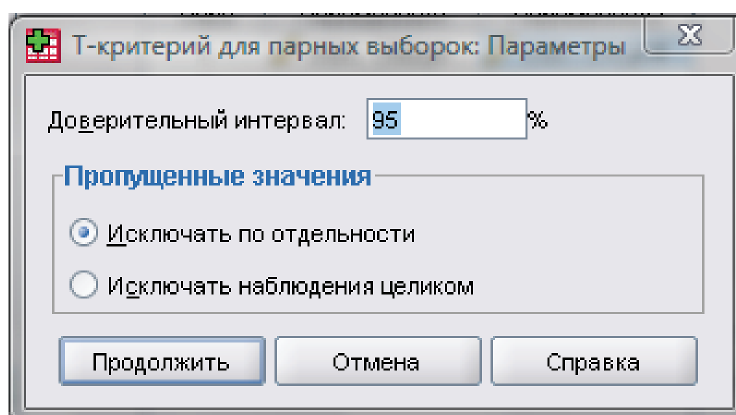


Рис. 7.12. Задача параметров в диалоговом окне «Т-критерий для парных выборок: параметры»

После того, как мы задали необходимые параметры и запустили процедуру, окно вывода выдает три таблицы: «Статистики парных выборок», «Корреляции парных выборок», «Критерии парных выборок». Данные пер-

вой и третьей таблицы нам были рассмотрены выше, остановимся на таблице «Корреляции парных выборок». Она показывает исследователю меру линейной связи между двумя переменными. Значения коэффициента корреляции может варьироваться в диапазоне от -1 до 1, знак коэффициента указывает направление отношений, и его абсолютная величина указывает силу взаимосвязи.

Однофакторный дисперсионный анализ (ANOVA)

Когда мы рассматривали процедуру М средние, нами уже была рассмотрена процедура однофакторного дисперсионного анализа. В программе SPSS имеется возможность более глубокой реализации этой функции. Учитывая наличие большого количества критериев, не будем останавливаться на этом вопросе подробно. Рассмотрим лишь сущность этого метода и кратко охарактеризуем используемые в нем критерии.

Процедура Однофакторный дисперсионный анализ (ANOVA) выполняет однофакторный дисперсионный анализ для количественной зависимой переменной по единственной факторной (независимой) переменной. Дисперсионный анализ используется для проверки гипотезы о равенстве нескольких средних значений, соответствующих различным группам или уровням факторной переменной. Этот метод является расширением двухвыборочного Т-критерия.

В дополнение к выявлению различий между средними значениями можно узнать, какие именно групповые средние значения различаются. Есть два типа критериев для сравнения средних значений: априорные контрасты и апостериорные критерии. Априорные контрасты – это критерии, которые применяются до проведения эксперимента, апостериорные же критерии применяются после проведения эксперимента. В SPSS имеется возможность осуществлять проверку наличия трендов по уровням (категориям).

В разделе справки SPSS приводится следующий пример возможности реализации однофакторного дисперсионного анализа. Пончики впитывают различное количество жира в процессе их приготовления. В эксперименте используются три типа жиров: арахисовое масло, кукурузное масло и свиное сало. Арахисовое и кукурузное масло являются ненасыщенными жирами, а топленое сало – насыщенным жиром. Выясняя, зависит ли количество расходуемого жира от типа используемого жира, можно выбрать априорный контраст, позволяющий выяснить, различаются ли количества впитывающегося жира для насыщенных и ненасыщенных жиров.

Данные, используемые в этом виде анализа, должны быть следующими: факторные переменные должны быть целочисленными, а зависимая переменная – количественной (измерена по крайней мере в интервальной шкале).

В соответствии с гипотезой однофакторного дисперсионного анализа, каждая группа является независимой случайной выборкой из нормального распределения. Дисперсионный анализ робастен (устойчив) к отклонениям от нормальности, однако данные должны быть симметричны. Группы должны выбираться из совокупностей с одинаковыми дисперсиями. Для проверки последнего предположения должен использоваться критерий Ливиня однородности дисперсий.

Диалоговое окно процедуры однофакторный дисперсионный анализ выглядит так:

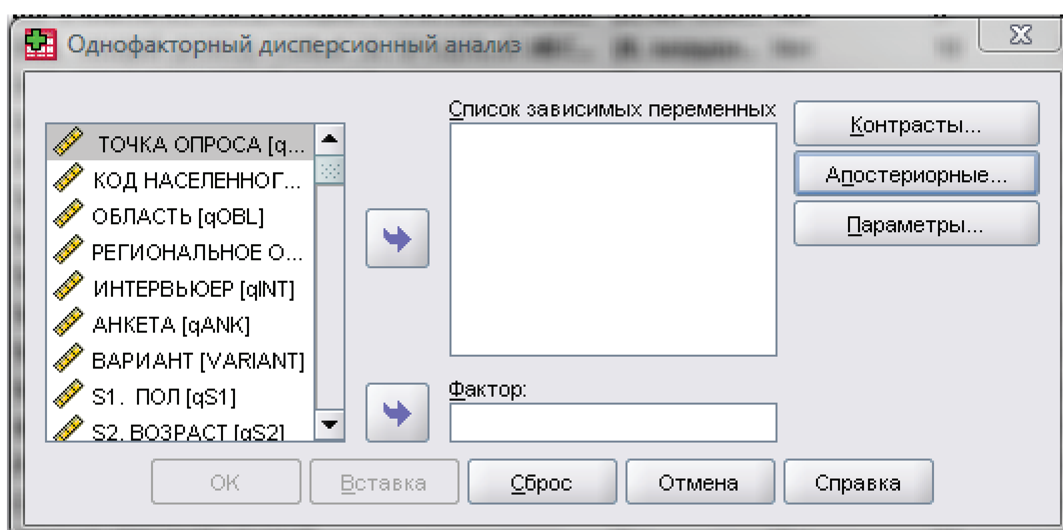


Рис. 7.13 Диалоговое окно «Однофакторный дисперсионный анализ»

При нажатии кнопки «Контрасты» появляется диалоговое окно следующего вида:

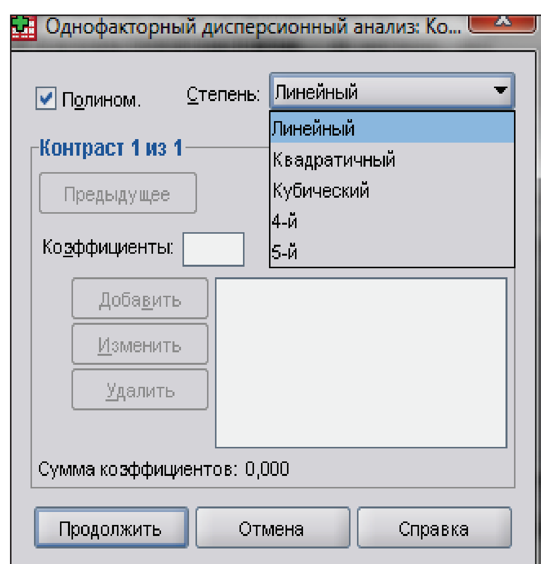


Рис. 7.14 Диалоговое окно «Однофакторный дисперсионный анализ: Коэффициенты»

Здесь можно разделить межгрупповые суммы квадратов на трендовые компоненты или задать априорные контрасты.

Раздел «Полиноминальный» расчленяет межгрупповые суммы квадратов на трендовые компоненты. Можно выполнить проверку на наличие тренда зависимой переменной по упорядоченным уровням факторной переменной. Например, можно проверить наличие линейного тренда (возрастающего или убывающего) заработной платы по упорядоченным уровням переменной, характеризующей служебное положение или уровень образования. Также можно проверить наличие квадратичного, кубического тренда, а также трендов 4-й и 5-й степеней.

Раздел коэффициенты содержит задаваемые пользователем априорные контрасты, которые будут проверяться при помощи Т-критерия. Здесь можно ввести значение коэффициента для каждой группы (уровня, категории) факторной переменной, а после ввода очередного значения необходимо щелкнуть мышью по кнопке «Добавить». Каждое новое значение будет добавлено в конец списка коэффициентов. Задать дополнительные наборы контрастов можно, щелкая по кнопке «Следующее». Для перехода от одного набора контрастов к другому используются кнопки «Следующее», «Предыдущее».

Порядок ввода коэффициентов важен, так как он соответствует возрастающему порядку значений категорий факторной переменной. Первый коэффициент в списке соответствует наименьшему значению факторной переменной, а последний — наибольшему. Например, если факторная переменная имеет шесть категорий, коэффициенты –1, 0, 0, 0, 0.5, 0.5 сопоставляют первую группу с пятой и шестой группами. В большинстве случаев сумма коэффициентов должна быть равна нулю. Наборы с ненулевой суммой также могут быть использованы, однако в этом случае появится предупреждающее сообщение.

При нажатии на кнопку «Апостериорные» появляется диалоговое окно следующего вида (рис. 7.15).

Установив, что различия средних значений существуют, с помощью апостериорных критериев размаха и парных множественных сравнений можно выяснить, какие именно средние различаются. Критерии размаха выявляют однородные подмножества средних, не различающихся между собой. Парные множественные сравнения проверяют разности между каждой парой средних значений и выдают матрицу, в которой звездочками обозначены групповые средние, значимо различающиеся на уровне альфа, равном 0,05.

При равенстве дисперсий используются следующие критерии:

НЗР. Использует Т-критерии для проведения всех парных сравнений групповых средних. Поправка для уровня ошибки на множественность сравнений не делается.

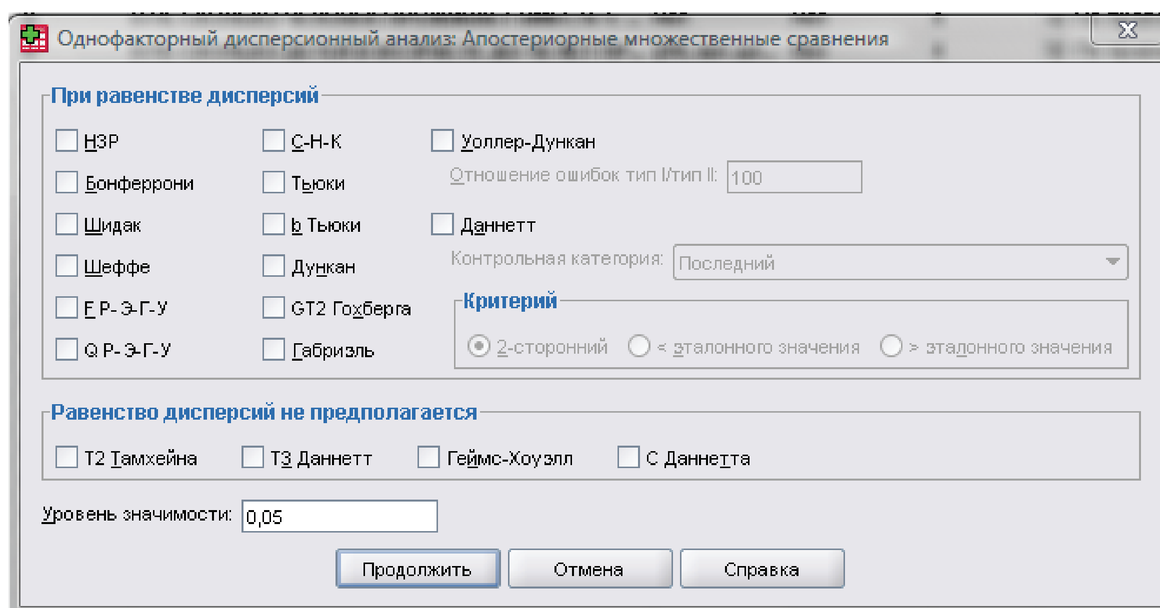


Рис. 7.15 Диалоговое окно «Однофакторный дисперсионный анализ: Апостериорные множественные сравнения»

Бонферрони. При проведении парных сравнений групповых средних используются T-критерии, но для управления общим уровнем ошибки по уровню ошибки каждой проверки вероятность ошибочного решения делится на общее число проверок. Доверительные интервалы и уровень значимости корректируются так, чтобы учесть проводимые множественные сравнения.

Шидак. Критерий множественных попарных сравнений, основанный на T-статистике. Критерий Шидака изменяет величину уровня значимости в соответствии с числом множественных сравнений и обеспечивает более узкие границы, чем критерий Бонферрони.

Шеффе. Производит одновременные сравнения совместных пар для всех возможных комбинаций пар средних. Использует выборочное F-распределение. Может применяться для проверки всех возможных линейных комбинаций групповых средних, а не только для парных сравнений.

Р-Э-Г-У F. Шаговая процедура множественных сравнений Райана-Эйнота-Габриэля-Уэлша, основанная на F-критерии.

Р-Э-Г-У Q. Шаговая процедура множественных сравнений Райана-Эйнота-Габриэля-Уэлша, основанная на стьюдентизированном размахе.

С-Н-К. В соответствии с критерием Стьюдента-Ньюмена-Келса выполняются все попарные сравнения средних, используя распределение стьюдентизированного размаха. Если объемы выборок одинаковы, с помощью шаговой процедуры сравниваются также пары средних в однородных подмножествах. Средние упорядочиваются по убыванию, а вначале проверяются наибольшие разности.

Тьюки. Использует статистику стьюдентизированного размаха для проведения всех парных сравнений между группами. Подгоняет уровень ошибки эксперимента к уровню ошибки совокупности всех парных сравнений.

b Тьюки. Для проведения парных сравнений между группами используется распределение стьюдентизированного размаха. Критической статистикой служит среднее из критических статистик двух критериев: достоверно значимой разности Тьюки и Стьюдента-Ньюмена-Келса.

Дункан. Выполняются парные сравнения с использованием шагового порядка сравнений, как и в критерии Стьюдента-Ньюмена-Келса, но устанавливается защитный уровень доли ошибок для набора проверок, а не для доли ошибок отдельных проверок. Основан на статистике стьюдентизированного размаха.

GT2 Гохберга. Критерий множественных сравнений и размахов, использующий стьюдентизированный максимум модуля. Аналогичен критерию достоверно значимой разности Тьюки.

Габриэль. Критерий парных сравнений, использующий стьюдентизированный максимум модуля, обычно более мощный, чем критерий Гохберга GT2, когда размеры ячеек не равны. Критерий Габриэля может стать либеральным, когда размеры ячеек сильно различаются.

Уоллер-Дункан. Процедура множественных сравнений, основанная на Т-статистике; использует байесовский подход.

Даннетт. Т-критерий множественных парных сравнений, который сравнивает средние по группам (уровням фактора) с одним контрольным средним. Последняя категория по умолчанию рассматривается как контрольная. Как вариант можно выбрать первую категорию. Двухсторонний проверяет, что среднее на любом из уровней (за исключением контрольной категории) фактора не равно среднему для контрольной категории. «< Эталона» проверяет, не окажется ли среднее на каком-либо из уровней фактора меньше, чем в контрольной категории. «>Эталона» проверяет, не окажется ли среднее на каком-либо из уровней фактора больше, чем в контрольной категории.

В случае, если **равенство дисперсий не предполагается**, используются критерии множественных сравнений:

Т2 Тамхейна. Консервативный критерий парных сравнений, основанный на Т-критерии.

Т3 Даннетт. Критерий парных сравнений, основанный на стьюдентизированном максимуме модуля.

Геймс-Хоуэлл. Критерий парных сравнений, иногда являющийся либеральным.

С Даннетта. Критерий парных сравнений, основанный на стьюдентизированном размахе.

Открывая диалоговое окно «Параметры», появляется соответствующее диалоговое окно (рис. 7.16).

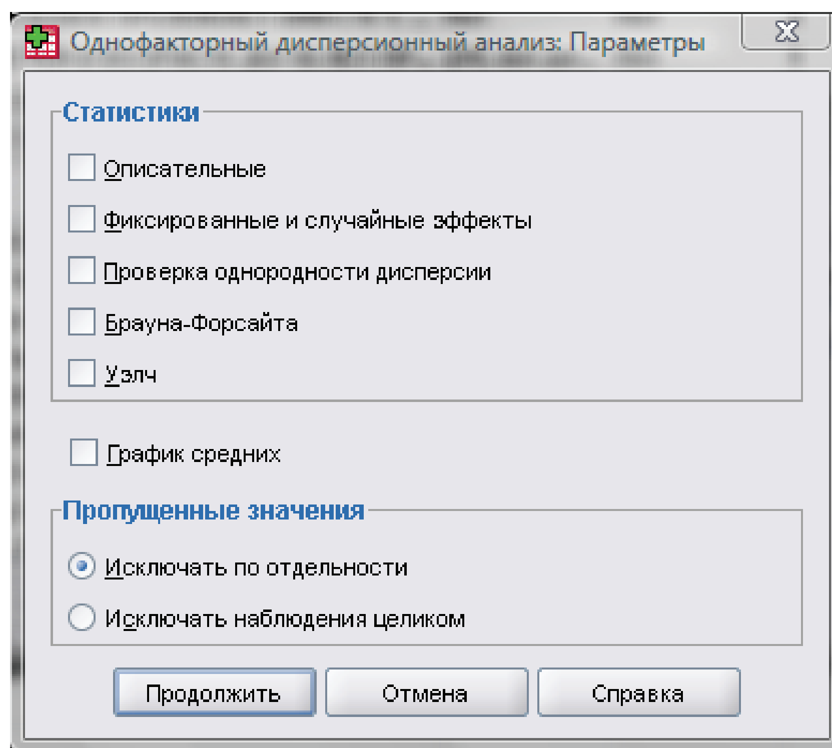


Рис. 7.16. Диалоговое окно «Однофакторный дисперсионный анализ: Параметры»

Здесь можно выбрать одну или несколько возможностей. В разделе **«Статистики»**:

Описательные. Для каждой зависимой переменной и каждой группы вычисляются: количество наблюдений, среднее значение, стандартное отклонение, стандартная ошибка среднего значения, минимум, максимум и доверительные интервалы в 95%.

Фиксированные и случайные эффекты. Выводит стандартное отклонение, стандартную ошибку и доверительный интервал в 95% для модели с фиксированными эффектами, а также стандартную ошибку, доверительный интервал в 95% и оценку межкомпонентной дисперсии для модели со случайными эффектами.

Проверка однородности дисперсии. Вычисляется статистика Ливиня для проверки равенства дисперсий групп. Этот критерий не требует предположения о нормальности.

Брауна-Форсайта. Вычисляется статистика Брауна-Форсайта для проверки равенства дисперсий групп. Эта статистика предпочтительнее F-статистики в случае, когда требование равенства дисперсий не выполняется.

Уэлч. Вычисляется статистика Уэлча для проверки равенства дисперсий групп. Эта статистика предпочтительнее F-статистики в случае, когда требование равенства дисперсий не выполняется.

«График средних» выводит график, изображающий средние подгрупп (средние для всех групп, заданных значениями факторной переменной).

Группа параметров **«Пропущенные значения»** позволяет управлять обработкой пропущенных значений:

Исключать по отдельности. Наблюдение с пропущенным значением зависимой или факторной переменной не используется в анализе. Не будут также использоваться наблюдения со значениями вне заданного диапазона факторной переменной.

Исключать целиком. Наблюдения с пропущенными значениями для факторной переменной или для любой из зависимых переменных, в списке зависимых переменных главного диалогового окна, не рассматриваются. Если не задано несколько независимых переменных, выбор этого параметра не играет роли.

Вопросы и задания

1. Дайте общую характеристику процедуры сравнения средних и ее возможностей.
2. Какие виды анализа в рамках общей группы «сравнение средних» реализованы в SPSS?
3. Приведите примеры использования различных процедур сравнения средних в практике анализа данных социологических исследований.
4. Самостоятельно осуществите процедуру сравнения средних, используя данные ЕАЭСД.

Список литературы

1. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цёфель. – СПб.: ДиаСофтЮП, 2005. – 608 с.

2. Венецкий, И. Г. Вариационные ряды и их характеристики / И. Г. Венецкий. – М.: Статистика, 1970. – 159 с.

3. Елисеева, И. И. Группировка, корреляция, распознавание образов / И. И. Елисеева, В. О. Рукавишников. – М.: Статистика, 1977. 144 с.

4. Крыштановский, А. О. Анализ социологических данных / А.О. Крыштановский. – М.: Изд-во «ГУ ВШЭ», 2007. – 281 с.

5. Турундаевский, В. Б. Многомерный статистический анализ в экономических задачах. Компьютерное моделирование в SPSS / В. Б. Турундаевский, И. В. Орлова, Н. А. Концевая. – М.: Изд-во «Вузовский учебник», 2009. – 320 с.

Лекция 8. Кластерный анализ

Классификация и разбиение объектов на однородные группы является важной процедурой в социологических исследованиях. Выделить в исследуемом массиве наиболее схожие между собой объекты и объединить их в группы (кластеры) позволяет кластерный анализ. Применение этого метода в анализе социологических данных становится все популярнее, так как, во-первых, классификация вообще является фундаментальным научным принципом, а во-вторых, с развитием и распространением вычислительных машин и соответствующего программного обеспечения выполнение громоздких и трудновыполнимых математических расчетов существенно облегчается.

Если говорить кратко, суть кластерного метода состоит в том, что объекты группируются в кластеры исходя из вычисленных математически расстояний между ними. Наблюдения, имеющие между собой наименьшие расстояния по определенному набору признаков, попадают в один кластер. Между объектами в разных кластерах расстояние, как правило, больше, чем между объектами в одном кластере.

Выполнение процедуры кластеризации данных с помощью SPSS, являющегося современным пакетом статистических программ, не столь сложно. При всех возможных различиях, которые могут появиться из-за разности проблематики и специфики параметров конкретного исследования, при проведении кластерного анализа обычно необходимы такие этапы: 1) отбор выборки объектов для кластеризации; 2) выбор признаков (переменных) для кластеризации; 3) вычисление меры сходства между объектами по выбранным признакам; 4) применение кластерного анализа для создания сходных групп объектов; 5) проверка достоверности полученной кластерной модели. Эта последовательность не зависит от применяемых алгоритмов кластеризации.

В пакете SPSS заложена возможность применения двух основных алгоритмов для кластеризации данных. Первый алгоритм – агломеративный иерархический, второй – итеративный по принципу k-средних. Считается, что применение первого – иерархического алгоритма кластеризации – предпочтительнее, поскольку он более чувствительно сортирует объекты наблюдения. Основная проблема с его применением возникает в случае чрезвычайно большого числа наблюдений или при учете множества признаков при кластеризации. В этом случае обычно применяется итеративный алгоритм.

Начинающему исследователю нужно иметь в виду, что полученные разными способами кластерные модели могут различаться как по количеству выделенных кластеров, так и по их наполнению. Это происходит из-за применения различных правил отбора наблюдений и использования различных статистических мер (евклидова расстояния, хи-квадрата или других).

В этой лекции мы рассмотрим на примерах применение обоих алгоритмов кластерного анализа с помощью SPSS. Однако прежде чем проиллюстрировать применение указанных алгоритмов, отметим общие моменты, касающиеся первых этапов выполнения кластерного метода — отбора объектов и выбора учитываемых признаков. С отбором объектов более-менее понятно, так как это, скорее всего, будет вся выборка вашего исследования. Что касается отбора учитываемых признаков, то тут следует иметь в виду один принципиально важный момент. Дело в том, что при вычислениях расстояний между объектами в расчетах участвует абсолютное значение учитываемых признаков. В том случае, если градации шкалы у разных признаков различны, то есть опасность, что признаки с большей градацией будут оказывать определяющее влияние на построение кластерной модели. Поясним на примере. Представьте, что вы хотите построить кластерную модель для рабочих предприятия N при двух учитываемых признаках — пол и стаж работы на предприятии. Первый признак имеет всего две градации: 1 — мужской, 2 — женский. Второй признак — стаж работы — может иметь большой вариационный размах — от 1 года до нескольких десятков лет. В таком случае возможные значения, которые примет вторая переменная, математически намного превышают значения первой переменной, поэтому и кластерная модель будет больше подвержена влиянию второй переменной. Это делает актуальной проблему стандартизации данных, то есть приведение их к сходной размерности. В нашем примере можно решить вопрос следующим образом. Обе переменные приведем к виду дихотомических шкал с вариантами ответов: 1 — «признак присутствует» и 0 — «признак отсутствует». Переменная «пол» будет разделена на две переменные — «женщины» и «мужчины», а переменная «стаж работы» — на ряд переменных в зависимости от необходимой дробности. Например, можно ее разделить на пять переменных — «менее года», «от 1 до 5 лет», «от 6 до 10 лет», «от 11 до 20 лет», «более 20 лет». В этом случае перекодирование переменных позволит решить проблему несоразмерности данных. Пример матрицы с переменными до и после стандартизации см. в таблицах 8.1 и 8.2.

Таблица 8.1

Данные до стандартизации

Gend	Stag	Var3	Var4	Var5	Var6	Var7
1	5					
2	11					
2	15					
1	2					
2	3					

Таблица 8.2.

Данные после стандартизации

Gend-m	Gend-f	Stag-do1	Stag1-5	Stag6-10	Stag11-20	Stag-b20
1	0	0	1	0	0	0
0	1	0	0	1	0	0
0	1	0	0	0	1	0
1	0	0	1	0	0	0
0	1	0	1	0	0	0

В реальной практике проведения социологических исследований не всегда можно выйти из ситуации таким образом. В каждом случае необходимо принимать решение исходя из ситуации и специфики данных. Лучше, если все учитываемые при кластеризации переменные имеют одинаковые шкалы. В этом случае стандартизация не нужна. Поэтому, если вы заранее решили использовать кластерный анализ, планируйте соразмерность шкал или способ решения этой задачи.

Иерархический алгоритм кластеризации

Иерархический агломеративный алгоритм кластеризации данных подразумевает пошаговое сравнение всех наблюдений и объединение наиболее схожих объектов в группы (кластеры). Поэтому такой алгоритм называется агломеративным или объединяющим.

Для того чтобы воспользоваться этим алгоритмом кластеризации данных в SPSS, необходимо в меню «Анализ» в разделе «Классификация» выбрать пункт «Иерархическая кластеризация» (рис. 8.1). После выбора этого пункта меню перед вами появится диалоговое окно следующего вида (рис. 8.2).

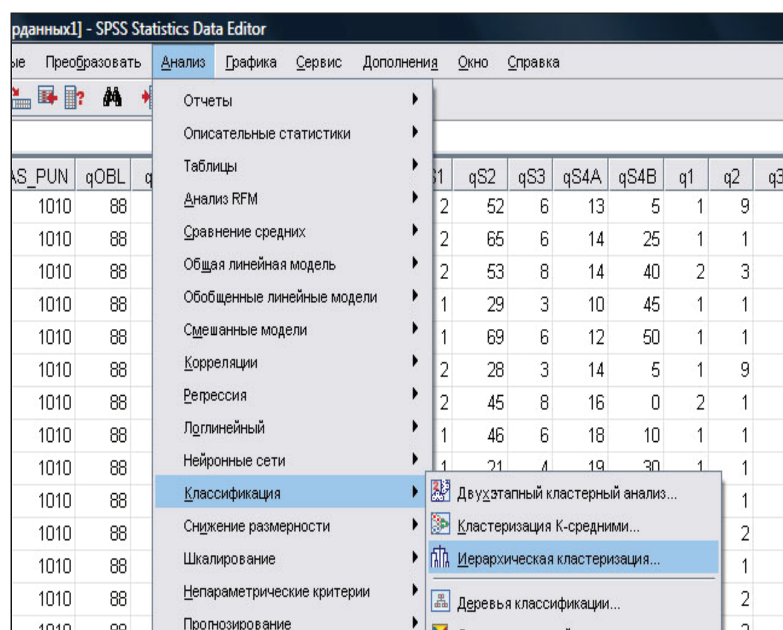


Рис. 8.1. Выбор пункта «Иерархическая кластеризация» в меню «Анализ: Классификация»

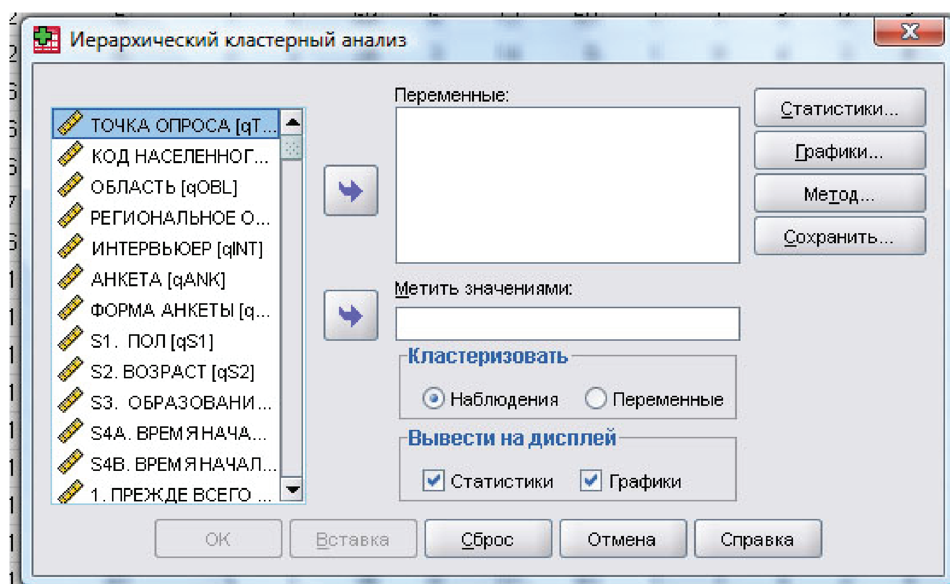


Рис. 8.2. Диалоговое окно «Иерархический кластерный анализ»

В левой части диалогового окна представлен перечень переменных, доступных для использования при кластеризации. В правой части – окна «Переменные» и «Метить значениями», а также ряд уже знакомых нам функциональных кнопок, назначение которых будет рассмотрено ниже.

Вспомним еще раз этапы кластерного анализа. Первый из них – отбор выборки объектов наблюдений. В нашем случае мы используем слишком большую базу объектов, поэтому, чтобы иерархический анализ получился наглядным (для учебного примера это необходимо), проведем случайную выборку 30 объектов с помощью процедуры «Отобрать наблюдения» в меню «Данные».

Далее необходимо выбрать признаки (переменные) для кластеризации наблюдений. Определим набор переменных для кластеризации, переместив их в окно «Переменные» в диалоговом окне «Иерархический кластерный анализ» (рис. 8.2). Для облегчения учебной задачи возьмем соразмерные данные (в этом случае нам не придется прибегать к стандартизации, о чем уже было сказано выше). В качестве примера попробуем построить кластерную модель с учетом пола респондентов и их ответов на вопрос о том, на кого они рассчитывают в трудных ситуациях с такими вариантами – «только на самого себя», «на своих родственников, друзей», «на помощь предприятия (организации), где работаю (работал)», «на помощь государства (органов соц. обеспечения)», «на помощь общественных организаций (профсоюз и т.п.)», «на благотворительную помощь», «на помощь церкви», «на другое», «затрудняюсь ответить». По каждому из ответов шкала дихотомическая, поэтому стандартизация не нужна. Переместим указанные переменные в окно для анализа переменных справа.

Третий этап – вычисление меры сходства объектов по выделенным признакам. Для этого воспользуемся кнопкой «Метод». В появившемся диалоговом окне (рис. 8.3) можно выбрать метод образования кластеров и меру расчета дистанции между наблюдениями, а также отметить, каким образом проводить стандартизацию данных и следует ли это делать.

SPSS предлагает исследователю широкий набор методов кластеризации. Здесь заложены возможности использования межгрупповых и внутригрупповых связей, методов ближайшего и дальнего соседа, центроидной и медианной кластеризации и, наконец, метода Варда.

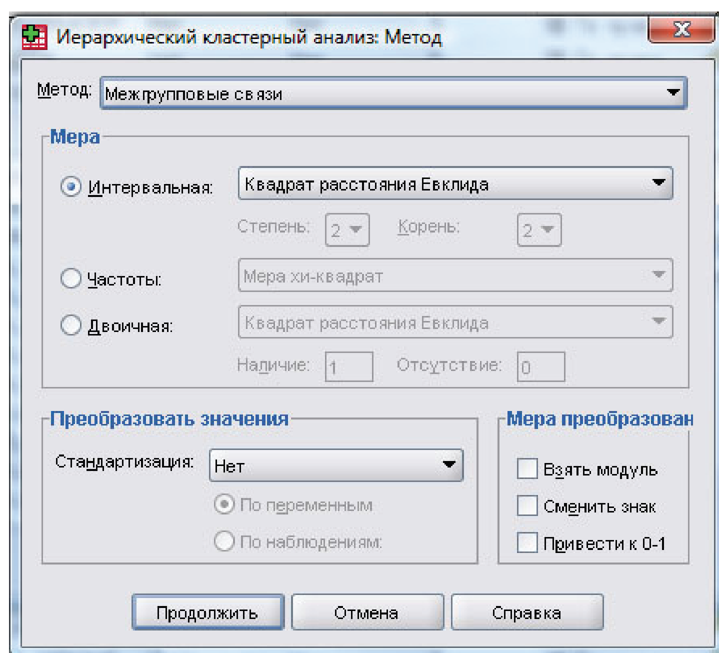


Рис. 8.3. Диалоговое окно «Иерархический кластерный анализ: Метод»

Метод межгрупповых связей основан на принципе, согласно которому расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Этот метод работает одинаково хорошо для формирования разных по форме кластеров и устанавливается программой по умолчанию.

Метод внутригрупповых связей идентичен предыдущему, однако при вычислениях учитываются парные связи внутри кластеров и размер соответствующих кластеров (то есть число объектов в них). Поэтому предлагаемый метод используется, когда предполагаются неравные размеры кластеров.

Метод ближайшего соседа при формировании кластеров использует расстояние между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. В результате происходит своеобразное «нанизывание» объектов на условную цепь. Таким образом, сформированные кластеры представляют собой продолговатые цепочки объектов.

Метод дальнего соседа расстояния между кластерами определяет наибольшим расстоянием между любыми двумя объектами в различных кластерах. Этот метод работает очень хорошо, когда объекты относятся к разным группам. В том случае, если кластеры имеют тенденцию к форме цепи (удлиненные по форме кластеры), то этот метод непригоден.

Метод центроидной кластеризации подразумевает вычисление расстояния между двумя кластерами как расстояния между их центрами тяжести. Центры тяжести кластеров являются усреднением значений объектов, относящихся к данному кластеру.

Метод медианной кластеризации похож на предыдущий. Его отличие состоит в том, что при вычислениях используются веса для учета разницы между размерами кластеров (то есть числами объектов в них). Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

Метод Варда при группировке объектов в кластеры использует методы дисперсионного анализа для оценки расстояний между кластерами, минимизируя сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге агломерации.

Выбор меры расстояния также является важным при построении кластерной модели. SPSS предлагает исследователю выбрать меры расстояния исходя из специфики анализируемых данных. Так, для интервальной шкалы можно использовать Евклидово расстояние или его квадрат, косинус, коэффициент корреляции Пирсона, коэффициент расстояния Чебышева, Блок (Манхэттенская мера), расстояние Минковского и пользовательскую меру, являющуюся разновидностью меры Минковского. Для частотных данных можно использовать хи- или фи-квадрат. Для двоичных данных – Евклидово расстояние или его квадрат, разность размеров и структур, дисперсию, разброс и ряд других мер. Разница применения различных мер расстояния настолько специфична в математическом смысле, что в учебной лекции мы не можем остановиться на этом подробно. Для изучения разницы указанных мер лучше обратиться к специальной литературе по математическому анализу. Здесь лишь отметим, что Евклидово расстояние и квадрат Евклидова расстояния являются достаточно универсальными мерами, широко применяемыми при кластерном анализе. Хи-квадрат, коэффициенты корреляции и размаха также вполне понятны интуитивно (мы их рассматривали в предыдущих лекциях).

Из предлагаемого набора методов кластеризации для социологических данных часто лучше всего подходит метод Варда. В нашем примере мы воспользуемся этим методом и выберем в качестве меры расстояния квадрат Евклидова расстояния, так как эта мера более наглядно разделяет кластеры

между собой. Теперь нажмем кнопку «Продолжить» и вернемся в диалоговое окно «Иерархический кластерный анализ».

Воспользовавшись кнопкой «Статистики» в появившемся диалоговом окне (рис. 8.4) выберем вывод порядка агломерации и матрицы близостей, а также укажем, что нам нужно не одно кластерное решение, а диапазон вариантов от 2 до 6 кластеров. Затем вновь нажмем кнопку «Продолжить».

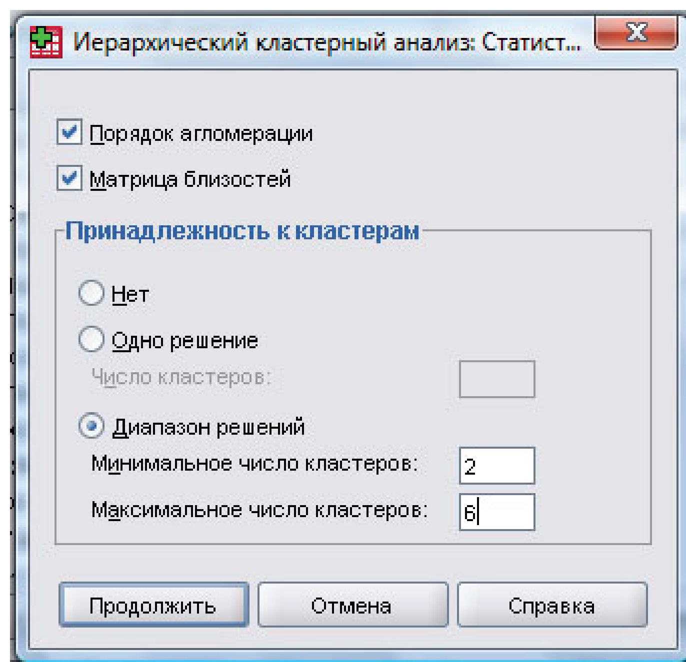


Рис. 8.4. Диалоговое окно «Иерархический кластерный анализ: Статистики»

Далее в окне «Графики» (рис. 8.5) отметим вывод древовидной диаграммы (дендрограммы) для всех без исключения кластеров. Это позволит представить последовательность объединения наблюдений в кластеры визуально в виде древовидной диаграммы (дендрограммы). Каждый шаг, на котором объединялась пара объектов, представляется ветвью этого дерева. На самом нижнем уровне все наблюдения независимы. Постепенно они объединяются в группы и, наконец, в одну большую группу на самом верхнем уровне. Затем в окне «Иерархический кластерный анализ» (рис. 8.2) нажмем «ОК».

Указанная последовательность шагов позволит получить в окне вывода программы следующие результаты, представленные на таблицах 8.3–8.6 и рисунках 8.6–8.7. Рассмотрим их по порядку.

Первой выводится небольшая по размеру таблица со сводкой обработки наблюдений (таблица 8.3), где представлена информация о том, сколько именно наблюдений обработано, сколько пропущено и какой метод кластеризации использовался. В нашем случае мы видим, что обработке подверглись 30 наблюдений, все из которых участвовали в формировании кластерной модели по методу Варда.

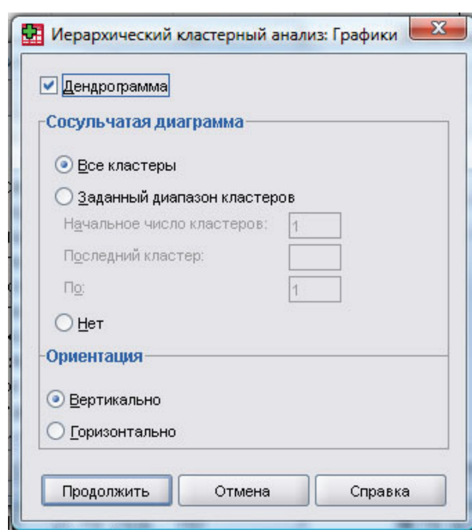


Рис. 8.5. Диалоговое окно «Иерархический кластерный анализ: Графики»

Таблица 8.3

Сводка обработки наблюдений

Наблюдения					
Валидные		Пропущенные		Всего	
N	Процент	N	Процент	N	Процент
30	100,0	0	,0	30	100,0

Следующая таблица (табл. 8.4) представляет собой матрицу близостей, где представлены расчеты квадрата Евклидова расстояния между всеми возможными парами объектов (в нашем случае это таблица размером 30x30). Ее вывод мы отметили для учебного примера. В реальной практике при значительном числе обрабатываемых наблюдений ее размеры слишком велики, чтобы их было целесообразно просматривать в таком виде. Поэтому в дальнейшем в окне «Иерархический кластерный анализ: Статистики» (рис. 8.4) следует убрать галочку напротив матрицы близостей.

Следующая таблица представляет собой распечатку порядка объединения наблюдений в кластеры (шагов агломерации) (см. таблицу 8.5). Из нее видно, что на первом шаге в общий кластер были объединены наблюдения под номерами 878 и 1441 (оба они – мужчины, рассчитывающие только на себя в трудной ситуации), затем участие этого кластера в дальнейшем объединении происходит на 11-м шаге, когда к наблюдению 878 присоединяется наблюдение 657 (также мужчина, рассчитывающий на себя) и т.д. Таким образом, данная таблица позволяет понять, как и в каком порядке происходило формирование кластеров. Естественно, что при числе наблюдений, значительно превышающем наш учебный массив, данная таблица также будет весьма громоздкой.

Таблица 8.4

Матрица близости

Наблю-дение	Квадраты Евклидовых расстояний																													
	117	146	178	181	209	231	275	333	336	337	370	391	431	470	499	601	657	671	878	895	1058	1088	1161	1205	1301	1347	1386	1425	1433	1441
117	,000	4,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	1,000	1,000	1,000	1,000	4,000	1,000	2,000	,000	2,000	1,000	1,000	,000	1,000	1,000	1,000	1,000	2,000	1,000	1,000	2,000
146	4,000	,000	3,000	3,000	3,000	3,000	3,000	2,000	3,000	4,000	3,000	3,000	3,000	3,000	2,000	3,000	2,000	4,000	2,000	3,000	3,000	4,000	3,000	3,000	3,000	3,000	2,000	3,000	3,000	2,000
178	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
181	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
209	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
231	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
275	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
333	2,000	2,000	3,000	1,000	1,000	3,000	1,000	,000	3,000	2,000	3,000	1,000	3,000	3,000	2,000	1,000	2,000	2,000	2,000	3,000	1,000	2,000	1,000	3,000	1,000	1,000	,000	1,000	3,000	2,000
336	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
337	,000	4,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	1,000	1,000	1,000	1,000	4,000	1,000	2,000	,000	2,000	1,000	1,000	,000	1,000	1,000	1,000	1,000	2,000	1,000	1,000	2,000
370	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
391	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
431	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
470	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
499	4,000	2,000	3,000	3,000	3,000	3,000	3,000	2,000	3,000	4,000	3,000	3,000	3,000	3,000	,000	3,000	2,000	4,000	2,000	3,000	3,000	4,000	3,000	3,000	3,000	3,000	2,000	3,000	3,000	2,000
601	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
657	2,000	2,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	2,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	2,000	,000	1,000	1,000	2,000	3,000	1,000	1,000	3,000	2,000	1,000	1,000	,000
671	,000	4,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	1,000	1,000	1,000	1,000	4,000	1,000	2,000	,000	2,000	1,000	1,000	,000	1,000	1,000	1,000	1,000	2,000	1,000	1,000	2,000
878	2,000	2,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	2,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	2,000	,000	1,000	1,000	2,000	3,000	1,000	1,000	3,000	2,000	1,000	1,000	,000
895	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
1058	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
1088	,000	4,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	1,000	1,000	1,000	1,000	4,000	1,000	2,000	,000	2,000	1,000	1,000	,000	1,000	1,000	1,000	1,000	2,000	1,000	1,000	2,000
1161	1,000	3,000	2,000	2,000	2,000	2,000	2,000	1,000	2,000	1,000	2,000	2,000	2,000	2,000	3,000	2,000	3,000	1,000	3,000	2,000	2,000	1,000	,000	2,000	2,000	,000	1,000	2,000	2,000	3,000
1205	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
1301	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
1347	1,000	3,000	2,000	2,000	2,000	2,000	2,000	1,000	2,000	1,000	2,000	2,000	2,000	2,000	3,000	2,000	3,000	1,000	3,000	2,000	2,000	1,000	,000	2,000	2,000	,000	1,000	2,000	2,000	3,000
1386	2,000	2,000	3,000	1,000	1,000	3,000	1,000	,000	3,000	2,000	3,000	1,000	3,000	3,000	2,000	1,000	2,000	2,000	2,000	3,000	1,000	2,000	1,000	3,000	1,000	1,000	,000	1,000	3,000	2,000
1425	1,000	3,000	2,000	,000	,000	2,000	,000	1,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	,000	1,000	1,000	1,000	2,000	,000	1,000	2,000	2,000	,000	2,000	1,000	,000	2,000	1,000
1433	1,000	3,000	,000	2,000	2,000	,000	2,000	3,000	,000	1,000	,000	2,000	,000	,000	3,000	2,000	1,000	1,000	1,000	,000	2,000	1,000	2,000	,000	2,000	2,000	3,000	2,000	,000	1,000
1441	2,000	2,000	1,000	1,000	1,000	1,000	1,000	2,000	1,000	2,000	1,000	1,000	1,000	1,000	2,000	1,000	,000	2,000	,000	1,000	1,000	2,000	3,000	1,000	1,000	3,000	2,000	1,000	1,000	,000

Это матрица различий

Таблица 8.5

Шаги агломерации

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	878	1441	,000	0	0	11
2	1205	1433	,000	0	0	7
3	1301	1425	,000	0	0	6
4	333	1386	,000	0	0	23
5	1161	1347	,000	0	0	23
6	181	1301	,000	0	3	13
7	178	1205	,000	0	2	14
8	671	1088	,000	0	0	12
9	601	1058	,000	0	0	13
10	470	895	,000	0	0	14
11	657	878	,000	0	1	25
12	117	671	,000	0	8	18
13	181	601	,000	6	9	20
14	178	470	,000	7	10	17
15	370	431	,000	0	0	17
16	275	391	,000	0	0	20
17	178	370	,000	14	15	21
18	117	337	,000	12	0	26
19	231	336	,000	0	0	21
20	181	275	,000	13	16	22
21	178	231	,000	17	19	28
22	181	209	,000	20	0	27
23	333	1161	1,000	4	5	26
24	146	499	2,000	0	0	25
25	146	657	3,800	24	11	28
26	117	333	6,300	18	23	27
27	117	181	9,550	26	22	29
28	146	178	13,536	25	21	29
29	117	146	21,700	27	28	0

Однако эта таблица жизненно необходима для определения оптимального числа кластеров. Считается, что оптимальной кластерной моделью является модель с таким числом кластеров, которое равно разности количества наблюдений (у нас – 30) и количества шагов агломерации, после которого значение коэффициента, то есть расстояния между двумя кластерами (в нашем случае квадрат Евклидова расстояния), увеличивается скачкообразно (у нас – 24). Попросту говоря, если не остановить кластеризацию на этом этапе, то в общий кластер будут объединены значения, которые отстоят друг от друга слишком далеко, то есть слишком разные наблюдения будут подвергнуты объединению в один кластер. Таким образом, мы выяснили, что оптимальной в нашем примере будет 6-кластерная модель.

Таблица 8.6

Принадлежность к кластерам

Наблюдение	6 кластеров	5 кластеров	4 кластеров	3 кластеров	2 кластера
117	1	1	1	1	1
146	2	2	2	2	2
178	3	3	3	3	2
181	4	4	4	1	1
209	4	4	4	1	1
231	3	3	3	3	2
275	4	4	4	1	1
333	5	5	1	1	1
336	3	3	3	3	2
337	1	1	1	1	1
370	3	3	3	3	2
391	4	4	4	1	1
431	3	3	3	3	2
470	3	3	3	3	2
499	2	2	2	2	2
601	4	4	4	1	1
657	6	2	2	2	2
671	1	1	1	1	1
878	6	2	2	2	2
895	3	3	3	3	2
1058	4	4	4	1	1
1088	1	1	1	1	1
1161	5	5	1	1	1
1205	3	3	3	3	2
1301	4	4	4	1	1
1347	5	5	1	1	1
1386	5	5	1	1	1
1425	4	4	4	1	1
1433	3	3	3	3	2
1441	6	2	2	2	2

Таблица 8.6 представляет собой таблицу принадлежности к кластерам, из которой видно, к какому из кластеров было отнесено каждое из наблюдений при построении различных кластерных моделей – от 2- до 6-кластерных (это мы указали при определении параметров кластеризации).

Благодаря этой таблице можно не только быстро понять, к какому кластеру было отнесено то или иное наблюдение, но и оценить, насколько успешна кластерная модель. Как известно, успешной кластерной моделью считается такая модель, в которой наблюдения более или менее устойчиво

относятся к определенному кластеру. Например, из таблицы 6 видно, что наблюдение 117 относится к кластеру 1 при всех апробированных моделях кластеризации, а наблюдение, например, 470 к кластеру 3. Вполне естественно, что при изменении общего числа кластеров некоторые наблюдения неизбежно будут перемещаться от кластера к кластеру – так происходит с наблюдениями под номерами 178, 181, 209, 275, 333 и другими. Однако, даже они перемещаются в разные кластеры не в каждой из моделей, на нескольких этапах продолжая оставаться в первоначально определенном для них кластере. Это говорит об определенной устойчивости построенных кластерных моделей, и поэтому следует признать эту работу успешной.

После всех таблиц выводятся графики – это сосульчатая диаграмма (рис. 8.6) и древовидная диаграмма (рис. 8.7).

Сосульчатая диаграмма (рис. 8.6) показывает, как происходил процесс объединения в кластеры. Внизу объединенные наблюдения отсутствуют, по мере чтения диаграммы вверх объединяемые наблюдения отмечаются столбиком в колонке между ними, тогда как различные кластеры указываются с помощью белых пробелов между ними. Однако такая диаграмма совершенно не наглядна, и ее вывод можно отменить на этапе определения параметров в диалоговом окне «Иерархический кластерный анализ: Графики» (рис. 8.5).

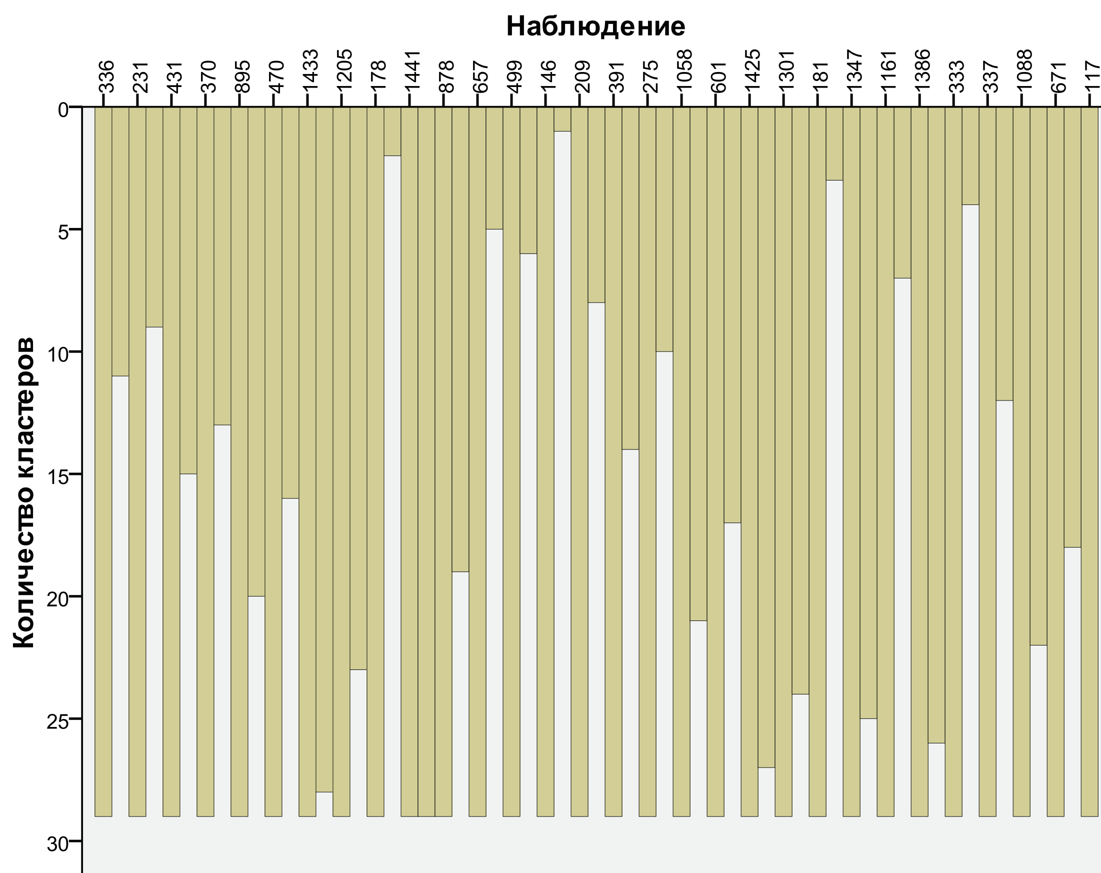


Рис. 8.6. Сосульчатая диаграмма

Куда более наглядна дендрогамма (рис. 8.7), на которой визуально представлен весь процесс объединения наблюдений в кластеры. Здесь можно увидеть, когда (при каком значении меры расстояния, выбранной нами для кластеризации) какое из наблюдений было объединено в кластеры. Здесь же можно увидеть, сколько именно кластеров будет и при каком пороговом значении меры расстояния. Так, если взять за основу квадрат Евклидова расстояния, равный 2, то получится 6-кластерная модель (она была признана оптимальной), если равный 10, то 3-кластерная модель и т.д. Если математически это можно понять из таблицы 8.5, то на дендрограмме это можно увидеть наглядно. Хотя, если увеличить число наблюдений до 100 или больше, то и дендрограмма перестанет быть наглядным средством представления процесса агломерации (объединения) наблюдений в кластеры.

```
* * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * *
 Dendrogram using Ward Method
 Rescaled Distance Cluster Combine

 C A S E 0 5 10 15 20 25
 Label Num +---------+---------+---------+---------+---------+

 878 -+
 1441 -+---------+
 657 -+ +-------------+
 146 -------+---+ |
 499 -------+ |
 1205 -+ |
 1433 -+ +-----------------------+
 178 -+ | |
 470 -+ | |
 895 -+ | |
 370 -+-----------------------+ |
 431 -+ |
 231 -+ |
 336 -+ |
 1301 -+ |
 1425 -+ |
 181 -+ |
 601 -+ |
 1058 -+ |
 275 -+-----------------+ |
 391 -+ | |
 209 -+ | |
 671 -+ +-----------------------------+
 1088 -+ |
 117 -+-------------+ |
 337 -+ +----+
 333 -+-----+ |
 1386 -+ +-------+
 1161 -+-----+
 1347 -+
```

Рис. 8.7. Древовидная диаграмма (дендрограмма)

Здесь мы не имеем возможности подробно обсуждать последний этап кластерного анализа – проверку надежности результатов. Мы оценили надежность в нашем примере косвенными признаками – устойчивостью кластеров, учетом начала скачкообразного объединения кластеров и т.д. В реальной исследовательской практике может возникнуть необходимость проверить надежность кластерной модели не только с помощью приблизительных инструментов, но и более точных статистических расчетов, таких как кофенетическая корреляция, многомерный дисперсионный анализ, повторная выборка, процедуры Монте-Карло и др. Об этих методах можно прочитать подробнее в специальной литературе.

Итеративный алгоритм кластеризации

Другой алгоритм кластеризации данных – итеративный алгоритм группировки изучаемых объектов по принципу К средних. Он применяется в случае, когда из-за большой выборки наблюдений или значительного числа учитываемых признаков провести иерархическую кластеризацию невозможно. При этом выбирается определенное количество максимально отдаленных друг от друга *k*-точек в *n*-мерном пространстве (*k* – количество выделяемых кластеров, а *n* – число учитываемых при группировке признаков). Эти точки на первом этапе рассматриваются как центры будущих кластеров. Затем вычисляется Евклидово расстояние каждого из оставшихся объектов (единиц наблюдения) до всех имеющихся *k*-центров.

Все единицы наблюдения распределяются по *k* кластерам в зависимости от того, к какому из кластерных центров они ближе. Далее в каждом кластере заново вычисляются координаты центра как средние покоординатные значения кластеров в *n*-мерном пространстве с учетом вновь включенных в кластер наблюдений. После чего все единицы наблюдения вновь перераспределяются. Итерации повторяются до тех пор, пока не будут найдены оптимальные (устойчивые) кластерные центры, то есть пока соотношение дисперсий внутри и вне кластера не примет максимальное значение. Последнее будет математическим свидетельством того, что объекты в кластерах существенно более схожи между собой, чем объекты из соседних кластеров.

Основной проблемой на данном этапе является определение исходного числа кластеров, так как использование итеративного метода не дает возможности определить число кластеров в процессе анализа, а требует определить его изначально. Эту весьма непростую задачу в нашем учебном примере мы решим, воспользовавшись результатами уже проведенного иерархического анализа. Попробуем провести кластеризацию объектов наблюдения по тем же признакам (пол респондента и ответы на вопрос о том, на кого или на что рассчитывает респондент в трудной ситуации), только

взяв за основу не 30 отобранных случайным образом объектов, а весь массив данных – 1601 наблюдение.

Итак, для того чтобы с помощью SPSS провести итеративный кластерный анализ, следует воспользоваться функцией «Кластеризация К средними» в разделе «Классификация» меню «Анализ» (рис. 8.8).

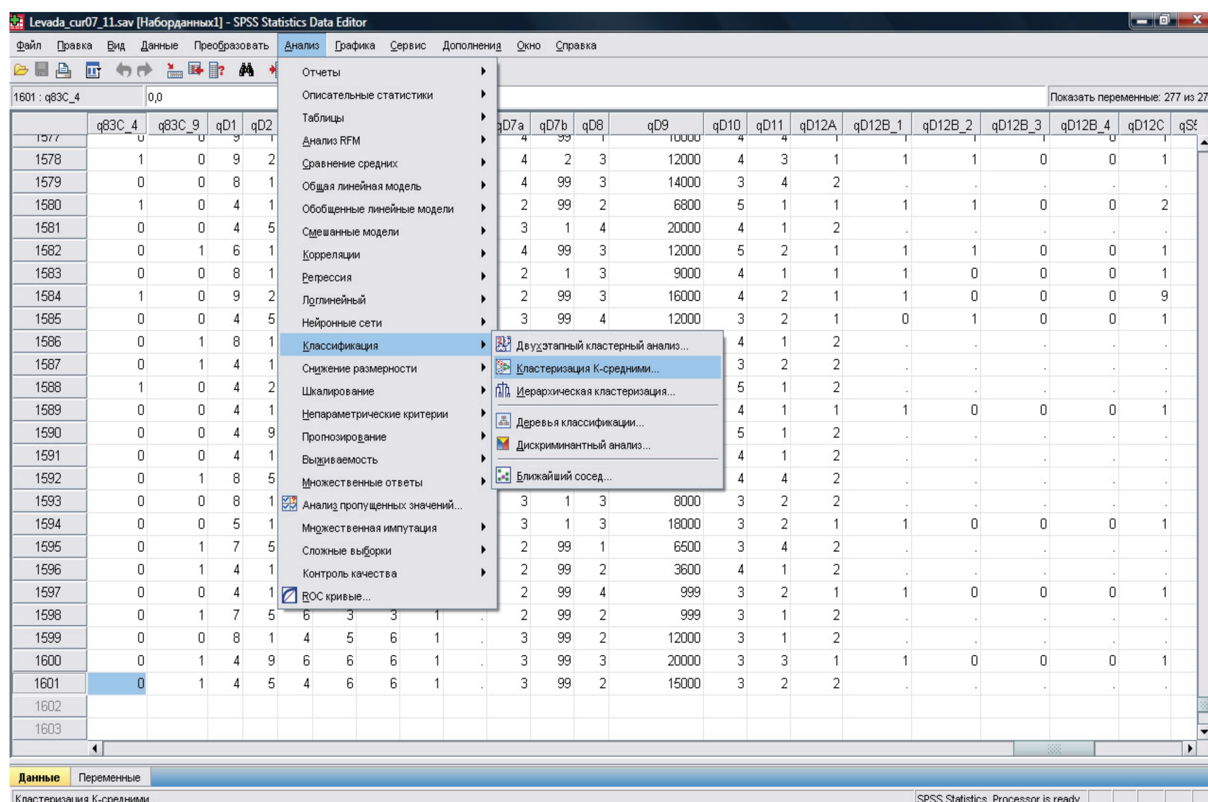


Рис. 8.8. Выбор процедуры итеративного кластерного анализа в меню «Анализ: Классификация»

После выбора соответствующего пункта меню перед вами появится диалоговое окно (рис. 8.9), в котором необходимо провести работу по определению параметров будущей кластеризации данных. Здесь необходимо задать перечень переменных для кластеризации, определить критерии и параметры совершения итераций.

Для начала переместим переменные qS1 и q48_1 – q48_9 в область анализируемых переменных. Ниже укажем число кластеров – 6. Если помните, мы получили это значение при проведении кластерного анализа по этим же переменным иерархическим способом для 30 наблюдений. Далее воспользуемся функциональной кнопкой «Итерации». В появившемся диалоговом окне (рис. 8.10) обозначим максимальное число возможных итераций числом 999 и нажмем «Продолжить».

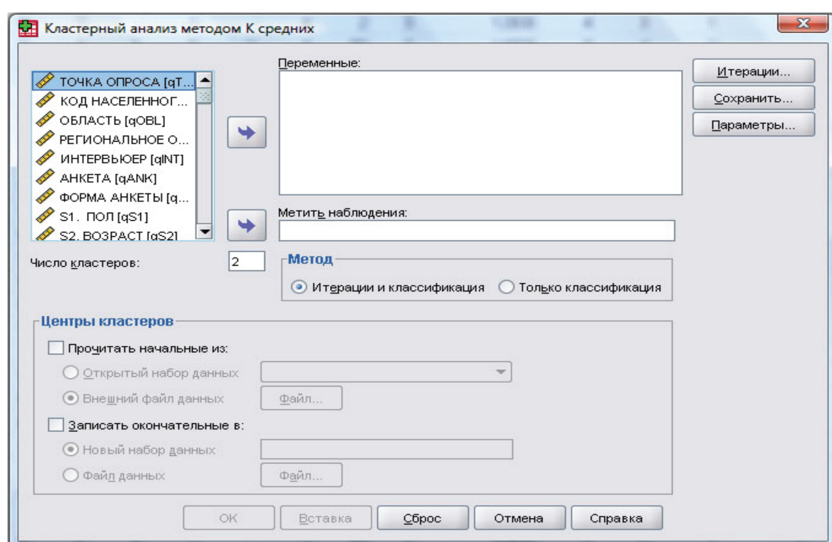


Рис. 8.9. Диалоговое окно «Кластерный анализ методом К средних»

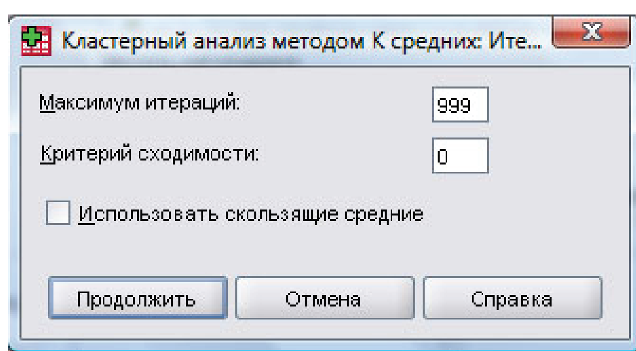


Рис. 8.10. Диалоговое окно «Кластерный анализ методом К средних: Итерации»

Затем в окне «Параметры» (рис. 8.11), появляющемся при нажатии одноименной кнопки, отметим вывод всех возможных статистик. Сделаем это только для учебного примера, так как в реальной практике это не всегда бывает нужно.

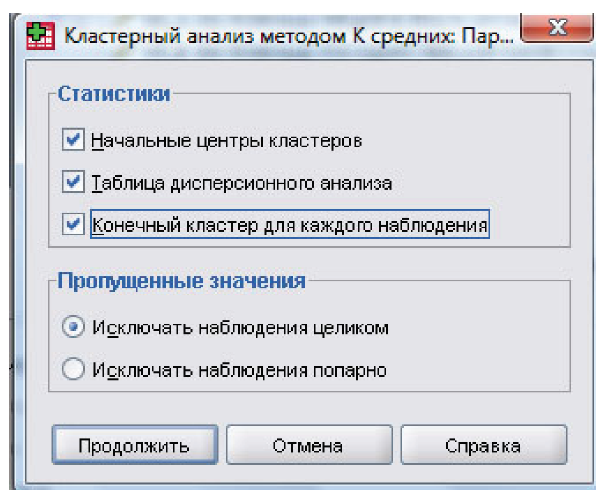


Рис. 8.11. Диалоговое окно «Кластерный анализ методом К средних: Параметры»

Если воспользоваться кнопкой «Сохранить» в окне «Кластерный анализ методом К средних» (рис. 8.9), то можно будет сохранить для каждого наблюдения привязку к определенному кластеру в виде отдельной переменной с номером кластера в качестве значения.

Для начала процедуры кластеризации нажмем «ОК» и получим результаты разбиения исследуемой совокупности на шесть групп респондентов. В результате данной процедуры мы получаем семь таблиц – это таблицы 8.7–8.13. Теперь по порядку объясним их содержание.

Первая выводимая таблица (таблица 8.7) содержит значение начальных центров кластеров для всех переменных, участвующих в обработке. Из нее мы видим, что первоначально первый кластер составили женщины, рассчитывающие в трудных ситуациях на себя, на помощь государства и общественных организаций. Во второй кластер попали женщины, рассчитывающие на помощь родственников, друзей и благотворителей. В третий кластер – женщины, рассчитывающие на себя и помощь церкви. В четвертый кластер – женщины, которые не знают, на что рассчитывать в трудной ситуации. В пятый кластер – мужчины, рассчитывающие на себя и помощь предприятия, на котором трудятся. И, наконец, в шестой кластер – мужчины, рассчитывающие на помощь родственников, друзей и государства. Но это не окончательное распределение, а только первоначально определенные центры кластеров, то есть наиболее не похожие друг на друга случаи.

Таблица 8.7

Начальные центры кластеров

	Кластер					
	1	2	3	4	5	6
S1. ПОЛ	2	2	2	2	1	1
48_1. ТОЛЬКО НА САМОГО СЕБЯ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	1	0	1	0	1	0
48_2. НА СВОИХ РОДСТВЕННИКОВ, ДРУЗЕЙ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	1	0	0	0	1
48_3. НА ПОМОЩЬ ПРЕДПРИЯТИЯ, ОРГАНИЗАЦИИ, ГДЕ РАБОТАЮ (РАБОТАЛ) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	1	0
48_4. НА ПОМОЩЬ ГОСУДАРСТВА (ОРГАНОВ СОЦИАЛЬНОГО ОБЕСПЕЧЕНИЯ) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	1	0	0	0	0	1

Окончание табл. 8.7

	Кластер					
	1	2	3	4	5	6
48_5. НА ПОМОЩЬ ОБЩЕСТВЕННЫХ ОРГАНИЗАЦИЙ (ПРОФСОЮЗ И Т.П.) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	1	0	0	0	0	0
48_6. НА БЛАГОТВОРИТЕЛЬНУЮ ПОМОЩЬ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	1	0	0	0	0
48_7. НА ПОМОЩЬ ЦЕРКВИ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	1	0	0	0
48_8. НА ДРУГОЕ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	0	0
48_9. ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0	0	0	1	0	0

Далее выводится история итераций (табл. 8.8). Здесь мы видим, что построение устойчивой 6-кластерной модели, при которой центры кластеров остаются на своих местах, завершено уже после четырех итераций, хотя мы и указали максимально возможное число – 999.

Таблица 8.8

История итераций

Итерация	Изменения центров кластеров					
	1	2	3	4	5	6
1	1,010	1,110	,953	,530	1,040	,809
2	,236	,036	,036	,354	,043	,172
3	,132	,000	,016	,300	,000	,052
4	,000	,000	,000	,000	,000	,000

a. Сходимость достигнута по критерию малой величины или отсутствия изменений в положении центров кластеров. Максимальное абсолютное изменение координаты любого кластера составляет ,000. Текущая итерация 4. Минимальное расстояние между начальными центрами 1,732.

Следующая таблица (табл. 8.9) иллюстрирует принадлежность каждого наблюдения к тому или иному кластеру и его расстояние от центра кластера. Ее мы приводим не полностью (только первые 10 строк), так как эта таблица включает в себя 1601 наблюдение, а соответственно 1601 строку.

Таблица 8.9

Принадлежность к кластерам

Номер наблюдения	Кластер	Расстояние
1	3	,013
2	2	,476
3	3	,989
4	5	,413
5	5	,588
6	2	,476
7	3	,013
8	5	,588
9	5	,588
10	5	,588

Таблица 8.10 аналогична таблице 8.7, только, в отличие от нее, содержит информацию не о начальных, а о конечных центрах кластеров.

Таблица 8.10

Конечные центры кластеров

	Кластер					
	1	2	3	4	5	6
S1. ПОЛ	2	2	2	1	1	1
48_1. ТОЛЬКО НА САМОГО СЕБЯ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	1	1	0	1	0
48_2. НА СВОИХ РОДСТВЕННИКОВ, ДРУЗЕЙ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	1	0	0	0	1
48_3. НА ПОМОЩЬ ПРЕДПРИЯТИЯ, ОРГАНИЗАЦИИ, ГДЕ РАБОТАЮ (РАБОТАЛ) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	0	0
48_4. НА ПОМОЩЬ ГОСУДАРСТВА (ОРГАНОВ СОЦИАЛЬНОГО ОБЕСПЕЧЕНИЯ) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	1	0	0	0	0	0
48_5. НА ПОМОЩЬ ОБЩЕСТВЕННЫХ ОРГАНИЗАЦИЙ (ПРОФСОЮЗ И Т.П.) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	0	0

Окончание табл. 8.10

	Кластер					
	1	2	3	4	5	6
48_6. НА БЛАГОТВОРИТЕЛЬНУЮ ПОМОЩЬ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	0	0
48_7. НА ПОМОЩЬ ЦЕРКВИ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	0	0
48_8. НА ДРУГОЕ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	0	0	0	0	0	0
48_9. ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0	0	0	0	0	0

Таблица 8.11 показывает Евклидово расстояние между полученными конечными центрами кластеров, то есть насколько отстоят друг от друга центры выделенных кластеров. Из таблицы мы видим, что все кластеры достаточно различаются между собой, так как их центры отстоят друг от друга на достаточном расстоянии. При этом наиболее близки друг к другу центры второго и третьего кластеров. Их близость очевидна и с логической точки зрения, так как второй кластер – это женщины, рассчитывающие на себя, родственников и друзей в трудной ситуации, а третий кластер – женщины, рассчитывающие только на себя. А наиболее отстоящие друг от друга кластеры – третий и шестой. В третий стремятся женщины, рассчитывающие на себя, а в шестой – мужчины, надеющиеся на родственников.

Таблица 8.11

Расстояния между конечными центрами кластеров

Кластер	1	2	3	4	5	6
1		1,456	1,254	1,169	1,412	1,546
2	1,456		1,107	1,334	1,253	1,131
3	1,254	1,107		1,229	1,082	1,734
4	1,169	1,334	1,229		1,287	1,215
5	1,412	1,253	1,082	1,287		1,163
6	1,546	1,131	1,734	1,215	1,163	

Таблица 8.12 содержит результаты одномерного дисперсионного анализа переменных, участвующих в кластеризации. Здесь для каждой переменной указаны значения средних квадратов, степеней свободы, значение F-критерия и вероятность (значимость). Благодаря этому этапу анализа можно выявить переменные, которые не оказывают никакого влияния на классификацию объектов наблюдения.

Последняя таблица содержит данные о числе наблюдений в каждом выделенном кластере (см. таблицу 8.13). Из нее мы видим, что самый большой кластер (пятый) содержит 536 наблюдений, а самый маленький (первый) 32 наблюдения. То есть женщин, которые рассчитывают на помощь государства в трудной ситуации, – меньше всего, а больше всего – мужчин, которые в трудные времена рассчитывают только на себя. Это легко объяснить и с точки зрения банальной логики, однако подтверждение этого вывода с помощью реализации кластерного анализа делает его статистически и эмпирически достоверным.

Таблица 8.12

ANOVA

	Кластер		Ошибка			
	Средний квадрат	ст.св.	Средний квадрат	ст.св.	F	Знч.
S1. ПОЛ	75,589	5	,012	1595	6432,012	,000
48_1. ТОЛЬКО НА САМОГО СЕБЯ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	39,074	5	,080	1595	486,342	,000
48_2. НА СВОИХ РОДСТВЕННИКОВ, ДРУЗЕЙ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	53,632	5	,081	1595	658,638	,000
48_3. НА ПОМОЩЬ ПРЕДПРИЯТИЯ, ОРГАНИЗАЦИИ, ГДЕ РАБОТАЮ (РАБОТАЛ) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	,359	5	,020	1595	17,628	,000
48_4. НА ПОМОЩЬ ГОСУДАРСТВА (ОРГАНОВ СОЦИАЛЬНОГО ОБЕСПЕЧЕНИЯ) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	6,219	5	,014	1595	432,275	,000
48_5. НА ПОМОЩЬ ОБЩЕСТВЕННЫХ ОРГАНИЗАЦИЙ (ПРОФСОЮЗ И Т.П.) – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	,113	5	,003	1595	33,255	,000
48_6. НА БЛАГОТВОРИТЕЛЬНУЮ ПОМОЩЬ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	,153	5	,008	1595	20,176	,000
48_7. НА ПОМОЩЬ ЦЕРКВИ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	,328	5	,013	1595	24,840	,000
48_8. НА ДРУГОЕ – РАССЧИТЫВАЮ В ТРУДНЫХ СИТУАЦИЯХ	,017	5	,002	1595	9,335	,000
48_9. ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	1,248	5	,007	1595	188,235	,000

Значения F-статистики следует использовать только как индикатор, ведь кластеры выбирались так, чтобы максимизировать расхождения между наблюдениями из разных кластеров. Наблюденные уровни значимости не скорректированы соответственно, и потому их нельзя применять для проверки гипотезы о равенстве средних значений кластеров.

Таблица 8.13

Число наблюдений в каждом кластере

Кластер	1	32,000
	2	485,000
	3	349,000
	4	45,000
	5	536,000
	6	154,000
Валидные		1601,000
Пропущенные значения		,000

На практике описание и подробная характеристика полученных кластеров требует серьезной работы, изучения разнообразных характеристик объектов для точного описания их типов, составляющих тот или иной класс (кластер). Представление кластеров в графическом виде с помощью диаграммы рассеяния имеет смысл, когда число учитываемых переменных не более трех, так как в противном случае мы не сможем представить графически n-мерное пространство.

Далее значения полученных кластеров можно сохранить в виде переменной и анализировать сходства и различия между различными показателями как кластерах, так и между ними.

Вопросы и задания

1. Для чего в социологии применяется кластерный анализ?
2. Последовательное выполнение каких этапов необходимо для проведения кластерного анализа? В чем их суть?
3. Какие процедуры кластерного анализа возможны в SPSS? В чем разница между ними?
4. Что такое дендрограмма? Что она дает в иерархическом кластерном анализе?
5. Как определяется оптимальное число кластеров при агломеративном алгоритме кластеризации?
6. Проинтерпретируйте полученные в лекции с помощью алгоритма итеративной кластеризации результаты кластерного анализа. Охарактеризуйте полученные кластеры и попытайтесь дать логическое объяснение их количественному наполнению.

Список литературы

1. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей/А. Бююль, П. Цёфель. — СПб.: ДиаСофтЮП, 2005. — 608 с.

2. Галицкая, Е. Г. Кластеры на факторах: как избежать распространенных ошибок?/Е. Г. Галицкая, Е. Б. Галицкий//Социология: 4 М. — 2006. — № 22. — С. 145–161.

3. Жамбю, М. Иерархический кластер-анализ и соответствия/М. Жамбю. — М.: Финансы и статистика, 1988. — 342 с.

4. Крыштановский, А. О. Анализ социологических данных/А. О. Крыштановский. — М.: Изд-во «ГУ ВШЭ», 2007. — 281 с.

5. Крыштановский, А. О. «Кластеры на факторах» — об одном распространенном заблуждении/А. О. Крыштановский//Социология: 4 М. — 2005. — № 21. — С. 172–187.

6. Мандель, И. Д. Кластерный анализ/И. Д. Мандель. — М.: Финансы и статистика, 1988. — 176 с.

7. Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — 215 с.

8. Черныш, М. Ф. Опыт применения кластерного анализа/М. Ф. Черныш//Социология: 4 М. — 2000. — № 12. — С. 129–141.

Лекция 9. Факторный анализ

В процессе проведения исследований социолог часто сталкивается с необходимостью структурировать и типологизировать данные. Попросту говоря, из массы разрозненных первичных данных нужно сложить, как из мозаики, общую картину состояния исследуемой совокупности (общества в целом или отдельных его групп). Вот тут на помощь социологу приходит такой метод математического анализа, как факторный анализ, который позволяет не только выявить наличие скрытых взаимосвязей между признаками, но и определить круг наиболее важных (влиятельных) переменных, благодаря чему появляется возможность сгруппировать исследуемую совокупность данных, объединив их в несколько достаточно крупных и значимых групп.

Основная идея этого метода анализа социологической информации заключена в том, что исследователь должен гипотетически предположить возможное количество факторов, влияющих на изучаемую совокупность данных, а затем, выделив эти факторы, попытаться их проинтерпретировать. В результате факторного анализа в заданное исследователем количество факторов будут объединены по степени сходства все исследуемые переменные. Это позволяет сократить число влияющих переменных до минимума и выделить объединенные факторы, оказывающие влияние на исследуемую совокупность. Основная сложность теперь будет состоять в интерпретации результатов факторного анализа. Успешный результат факторного анализа приближает исследователя к выявлению причинно-следственных связей и позволяет выделить более или менее однородные группы респондентов на основе выявленных факторов.

Теперь более детально остановимся на процедурах, необходимых для осуществления факторного анализа с помощью SPSS и интерпретации результатов факторного анализа.

Для иллюстрации возьмем пример исследования Института социологии РАН «Вопросник для взрослых», из которого нами будет взят блок вопросов, позволяющий составить социально-психологический портрет респондентов относительно их уверенности в себе и уровне самоуважения. В ходе опроса по четырехбалльной шкале (от «полностью согласен» до «совсем не согласен») фиксировалось отношение респондентов к 17 суждениям-мнениям:

Я не могу справиться со своими проблемами (С-1);

Иногда я чувствую, что мной помыкают в жизни (С-2);

Я мало могу влиять на то, что со мной происходит (С-3);

Я всегда могу выполнить задуманное (С-4);

Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни (С-5);

То, что со мной произойдет в будущем, во многом зависит от меня (С-6);

То, что я могу сделать, мало что изменит в моей жизни (С-7);

Я думаю, что я ничем не хуже других (С-8);

Я считаю, что у меня есть много хороших качеств (С-9);

В общем, мне кажется, что я неудачник (неудачница) (С-10);

Я могу все делать не хуже других (С-11);

Я думаю, что мне особенно нечем гордиться (С-12);

Я хорошо отношусь к самому (самой) себе (С-13);

В целом я удовлетворен (удовлетворена) собой (С-14);

Иногда я чувствую себя бесполезным (бесполезной) (С-15);

Я хотел (хотела) бы относиться к себе с большим уважением (С-16);

Иногда мне кажется, что я нехороший человек (С-17).

Теперь покажем на этом примере, что может дать социологу такой аналитический инструмент, как факторный анализ. Для начала работы выберем в меню программы пункт «Факторный анализ». Он содержится в разделе «Анализ» – «Снижение размерности» (рис. 9.1).

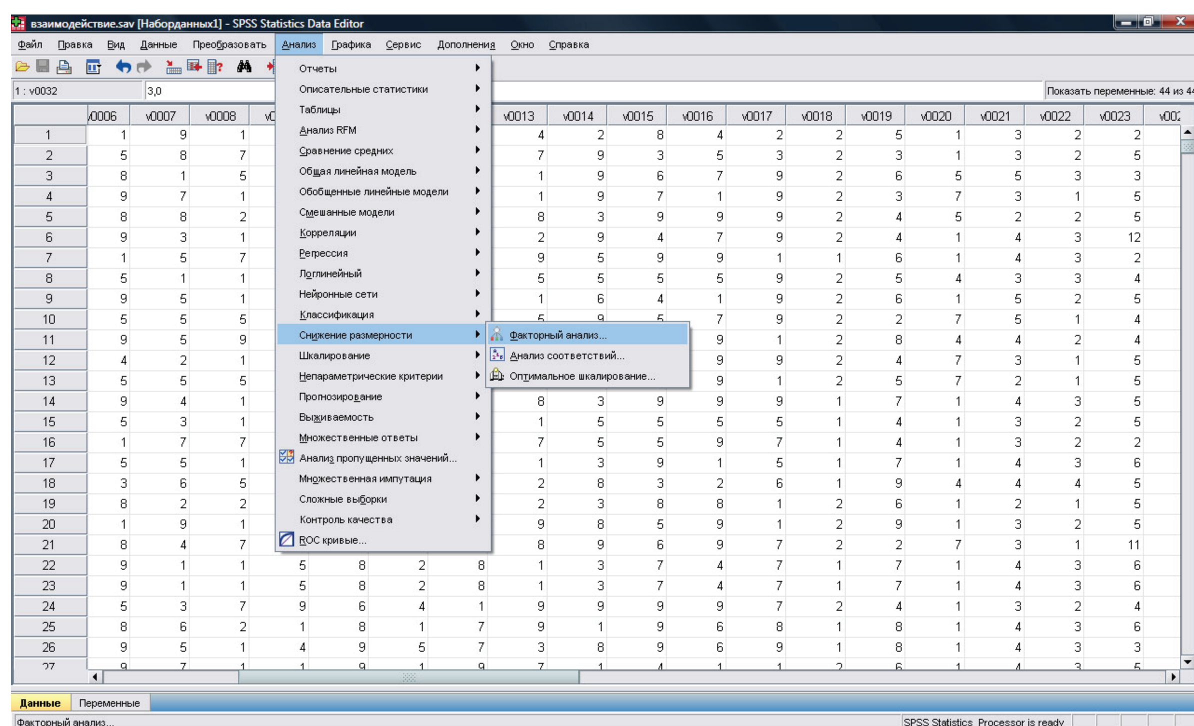


Рис. 9.1. Выбор пункта «Факторный анализ» в меню «Анализ: Снижение размерности»

Далее на экране появится диалоговое окно, похожее на подобные окна для других инструментов программы, но имеющее и свои отличия для процедуры факторного анализа (см. рис. 9.2).

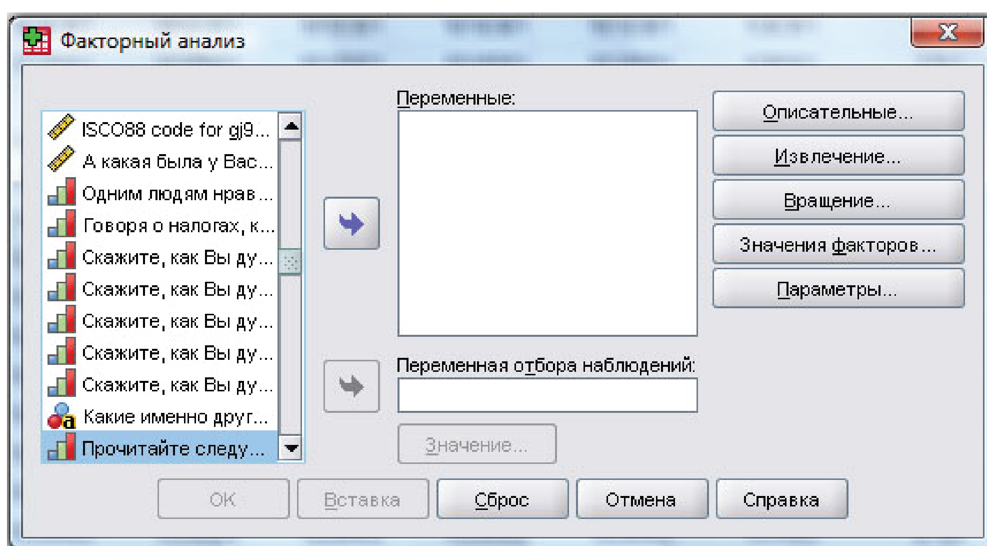


Рис. 9.2. Диалоговое окно «Факторный анализ»

В левой части данного диалогового окна, как и в других подобных окнах, виден список доступных для анализа переменных. В правой части – окна «Переменные» и «Переменная отбора наблюдений», а также ряд функциональных кнопок, назначение которых будет рассмотрено ниже. Перенесем рассматриваемые переменные в окно переменных для анализа и нажмем кнопку «Описательные…». В появившемся окне (рис. 9.3) можно выбрать вывод одномерных описательных статистик (среднее значение, стандартное отклонение и число значимых наблюдений) для преобразуемых в факторы переменных, начального решения (предварительные относительные дисперсии и процентные доли объясненной дисперсии), а также разные варианты корреляционных матриц. Выберем вывод и одномерных описательных статистик, и начального решения и нажмем кнопку «Продолжить».

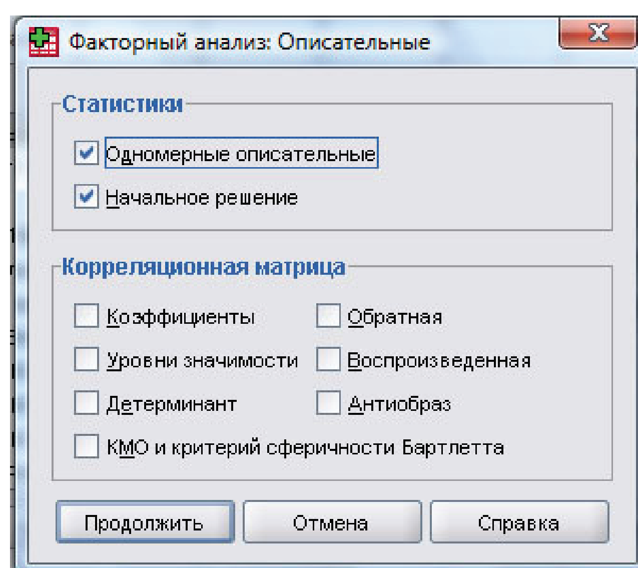


Рис. 9.3. Диалоговое окно «Факторный анализ: Описательные»

Далее обратимся к содержимому кнопки «Извлечение». Здесь можно выбрать метод снижения размерности, метод анализа, способ вывода на экран, количество выделяемых факторов и количество максимальных итераций (см. рис. 9.4).

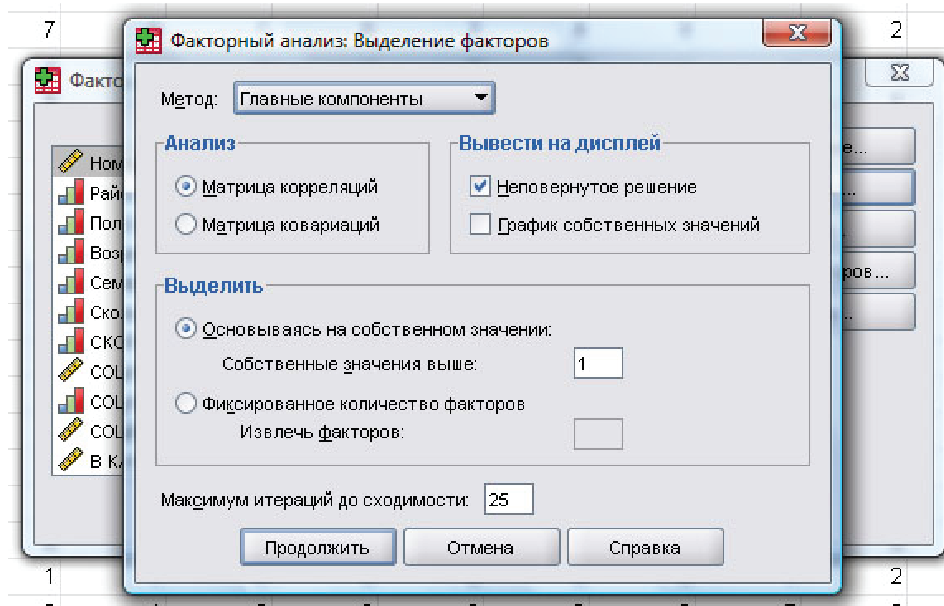


Рис. 9.4. Диалоговое окно «Факторный анализ: Выделение факторов»

Для проведения факторного анализа в SPSS исследователю предлагается несколько методов:

1. Метод главных компонент, основанный на последовательном поиске наиболее влиятельных факторов в порядке убывания их важности для объяснения зависимостей между переменными. В этом случае первый выделенный фактор будет объяснять наибольшую долю дисперсии признака, второй фактор – следующую наибольшую долю дисперсии и т.д.

2. Невзвешенный метод наименьших квадратов, минимизирующий сумму квадратов разностей между наблюденной и воспроизведенной корреляционной матрицами без учета диагоналей.

3. Обобщенный метод наименьших квадратов, похожий на предыдущий, но отличающийся от него тем, что корреляции взвешиваются величинами, обратными характерностям, так что переменные с высокой характерностью получают меньшие веса, чем переменные с низкой.

4. Метод максимального правдоподобия, при котором в качестве оценок параметров выбираются те, для которых наблюденная корреляционная матрица наиболее правдоподобна. При этом методе корреляции взвешиваются значениями, обратными к характерностям переменных.

5. Метод факторизации на главные оси, позволяет выделить факторы из исходной матрицы корреляций с квадратами коэффициентов множествен-

ных корреляций по диагонали в качестве начальных оценок общностей. В ходе факторного анализа с помощью этого метода последующие значения выделенных общностей замещают первоначальные до тех пор, пока не будет найдено наиболее подходящее решение.

Также можно воспользоваться методом альфа-факторизации и анализом Л. Гуттмана.

Для нашего примера мы воспользуемся одним из самых распространенных методов – методом главных компонент. Оставим все без изменений, добавив только отметку напротив графика собственных значений и нажмем кнопку «Продолжить». Остальные функциональные кнопки рассмотрим позднее. Нажав кнопку «ОК», мы получим следующий результат (см. таблицы 9.1–9.4, рис. 9.5).

Таблица 9.1

Описательные статистики

	Среднее	Стд. отклонение	Анализ N
Я не могу справиться со своими проблемами	2,29	1,195	10499
Иногда я чувствую, что мной помыкают в жизни	2,29	1,237	10499
Я мало могу влиять на то, что со мной происходит	2,46	1,316	10499
Я всегда могу выполнить задуманное	2,63	1,263	10499
Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни	2,48	1,209	10499
То, что со мной произойдет в будущем, во многом зависит от меня	2,89	1,181	10499
То, что я могу сделать, мало что изменит в моей жизни	2,57	1,260	10499
Я думаю, что я ничем не хуже других	3,23	1,016	10499
Я считаю, что у меня есть много хороших качеств	3,20	1,005	10499
В общем, мне кажется, что я неудачник (неудачница)	2,11	1,300	10499
Я могу все делать не хуже других	3,15	1,057	10499
Я думаю, что мне особенно нечем гордиться	2,36	1,271	10499
Я хорошо отношусь к самому (самой) себе	3,09	1,135	10499
В целом я удовлетворен (удовлетворена) собой	2,95	1,163	10499
Иногда я чувствую себя бесполезным (бесполезной)	2,37	1,250	10499
Я хотел (хотела) бы относиться к себе с большим уважением	2,94	1,187	10499
Иногда мне кажется, что я нехороший человек	2,11	1,210	10499

Таблица 9.2

Общности

Общности		
	Начальные	Извлеченные
Я не могу справиться со своими проблемами	1,000	,717
Иногда я чувствую, что мной помыкают в жизни	1,000	,668
Я мало могу влиять на то, что со мной происходит	1,000	,636

Окончание табл. 9.2

Я всегда могу выполнить задуманное	1,000	,566
Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни	1,000	,710
То, что со мной произойдет в будущем, во многом зависит от меня	1,000	,633
То, что я могу сделать, мало что изменит в моей жизни	1,000	,614
Я думаю, что я ничем не хуже других	1,000	,696
Я считаю, что у меня есть много хороших качеств	1,000	,704
В общем, мне кажется, что я неудачник (неудачница)	1,000	,673
Я могу все делать не хуже других	1,000	,670
Я думаю, что мне особенно нечем гордиться	1,000	,638
Я хорошо отношусь к самому (самой) себе	1,000	,647
В целом я удовлетворен (удовлетворена) собой	1,000	,660
Иногда я чувствую себя бесполезным (бесполезной)	1,000	,693
Я хотел (хотела) бы относиться к себе с большим уважением	1,000	,591
Иногда мне кажется, что я нехороший человек	1,000	,649
Метод выделения: Анализ главных компонент.		

Таблица 9.3

Полная объясненная дисперсия

Компонента	Начальные собственные значения			Суммы квадратов нагрузок извлечения		
	Итого	% Дисперсии	Кумулятивный %	Итого	% Дисперсии	Кумулятивный %
1	9,882	58,129	58,129	9,882	58,129	58,129
2	1,284	7,554	65,683	1,284	7,554	65,683
3	,584	3,434	69,117			
4	,561	3,299	72,417			
5	,502	2,950	75,367			
6	,443	2,606	77,973			
7	,424	2,493	80,466			
8	,400	2,350	82,816			
9	,379	2,230	85,046			
10	,362	2,131	87,177			
11	,349	2,052	89,230			
12	,335	1,970	91,200			
13	,325	1,914	93,113			
14	,313	1,842	94,956			
15	,294	1,729	96,684			
16	,290	1,705	98,390			
17	,274	1,610	100,000			

Метод выделения: Анализ главных компонент.

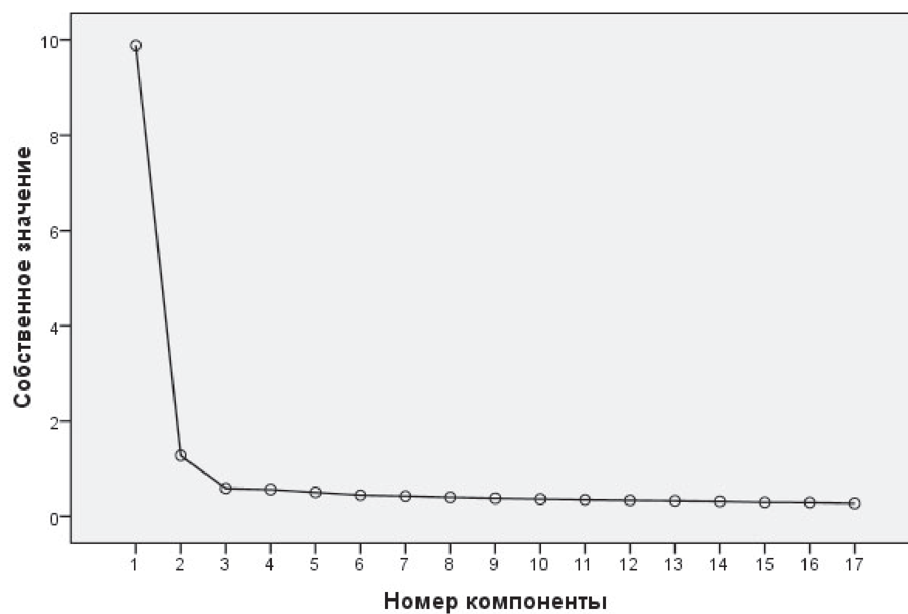


Рис. 9.5. График каменистой осыпи

Таблица 9.4

Матрица компонент

	Компонента	
	1	2
Я не могу справиться со своими проблемами	,788	-,310
Иногда я чувствую, что мной помыкают в жизни	,779	-,248
Я мало могу влиять на то, что со мной происходит	,736	-,306
Я всегда могу выполнить задуманное	,666	,350
Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни	,782	-,314
То, что со мной произойдет в будущем, во многом зависит от меня	,703	,374
То, что я могу сделать, мало что изменит в моей жизни	,748	-,232
Я думаю, что я ничем не хуже других	,773	,314
Я считаю, что у меня есть много хороших качеств	,787	,290
В общем, мне кажется, что я неудачник (неудачница)	,787	-,232
Я могу все делать не хуже других	,766	,289
Я думаю, что мне особенно нечем гордиться	,778	-,182
Я хорошо отношусь к самому (самой) себе	,744	,306
В целом я удовлетворен (удовлетворена) собой	,742	,331
Иногда я чувствую себя бесполезным (бесполезной)	,797	-,239
Я хотел (хотела) бы относиться к себе с большим уважением	,768	-,016
Иногда мне кажется, что я нехороший человек	,803	-,066

Метод выделения: Анализ методом главных компонент.

а. Извлеченных компонент: 2

Таблица 9.1 содержит простые описательные статистики для изучаемых переменных, как-то: среднее значение, стандартное отклонение и количество учитываемых (значимых) наблюдений.

Таблица 9.2 содержит информацию о том, какая часть дисперсии каждой из учитываемых в анализе переменных может быть объяснена предложенной факторной моделью. Например, мы видим, что отношение респондентов к суждению «Я не могу справиться со своими проблемами», согласие или несогласие с ним, объясняется данной факторной моделью на 71,7% (это максимальное значение), тогда как отношение к суждению «Иногда мне кажется, что я нехороший человек» только на 56,6%. Эта таблица может быть полезна для того, чтобы исключить слишком «непредсказуемые» переменные из анализа, тем самым повысив общую прогностическую успешность факторной модели. В нашем случае этого не требуется, так как все переменные объясняются более чем на 50%.

В таблице 9.3 видно, какую именно долю общей дисперсии может объяснить каждый из выделенных факторов в отдельности и вся построенная факторная модель в целом. В нашем случае предложенная компьютером на основе расчетов двухфакторная модель может объяснить 65,6% дисперсии. При этом первый фактор предсказывает 58% общей дисперсии, второй – 7,5%. Такая таблица может оказаться полезной для принятия исследователем окончательного решения о количестве выделяемых факторов, ведь не имеет смысла использовать фактор, который может объяснить слишком малую долю дисперсии. Такой фактор лучше убрать из общего числа факторов. Это упростит модель и позволит правильно ее интерпретировать.

График каменной осыпи (рис. 9.5) также позволяет определить количество эффективных факторов в модели. Впервые этот метод определения числа факторов предложил Р. Кеттел, рекомендовав найти на графике место, где дальнейшее увеличение числа факторов перестает быть условием повышения объяснительной способности факторной модели. В нашем случае видно, что трехфакторная модель может дать дополнительную информацию по сравнению с двухфакторной, но увеличение числа факторов до четырех или более – бессмысленно. Это видно и из таблицы 9.3, но на графике это представлено более наглядно.

Таблица 9.4 представляет собой матрицу факторных нагрузок, из которой видно, с каким из факторов наиболее сильно коррелирует та или иная переменная. В нашем примере отношение респондентов ко всем суждениям теснее всего связано с первым фактором, второй фактор значительно слабее. Такую модель сложно интерпретировать.

Учитывая то, что на основе таблицы полной объясненной дисперсии (табл. 9.3) и особенно графика каменистой осыпи (рис. 9.5) мы убедились в целесообразности изменения числа факторов, проведем небольшую коррек-

цию. Теперь в диалоговом окне «Факторный анализ: Выделение факторов» (рис. 9.4) зададим число факторов сами. Возьмем за основу трехфакторную модель с объяснительной способностью 69%. Для этого выберем в разделе «Выделить» диалогового окна «Факторный анализ: Выделение факторов» фиксированное количество факторов – 3 (см. рис. 9.6). Далее нажмем кнопки «Продолжить» и «ОК».

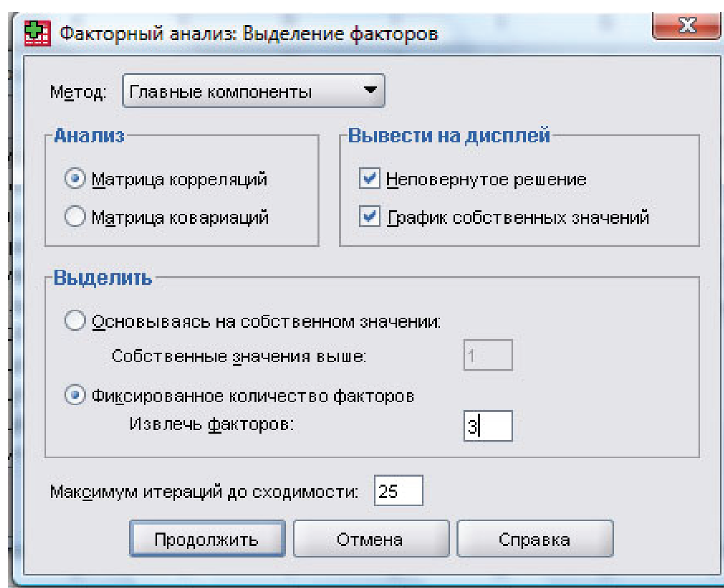


Рис. 9.6. Выбор фиксированного количества факторов в диалоговом окне «Факторный анализ: Выделение факторов»

Незначительные изменения в выводимых таблицах 9.1–9.3 нас не интересуют. Больший интерес представляет изменение матрицы факторных нагрузок – теперь в ней три фактора (см. табл. 9.5).

Таблица 9.5

Матрица компонент

	Компонента		
	1	2	3
Я не могу справиться со своими проблемами	,788	-,310	,184
Иногда я чувствую, что мной помыкают в жизни	,779	-,248	,129
Я мало могу влиять на то, что со мной происходит	,736	-,306	,297
Я всегда могу выполнить задуманное	,666	,350	,190
Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни	,782	-,314	,184
То, что со мной произойдет в будущем, во многом зависит от меня	,703	,374	,006
То, что я могу сделать, мало что изменит в моей жизни	,748	-,232	,122
Я думаю, что я ничем не хуже других	,773	,314	,123
Я считаю, что у меня есть много хороших качеств	,787	,290	,148

Окончание табл. 9.5

	Компонента		
	1	2	3
В общем, мне кажется, что я неудачник (неудачница)	,787	-,232	-,170
Я могу все делать не хуже других	,766	,289	,062
Я думаю, что мне особенно нечем гордиться	,778	-,182	-,275
Я хорошо отношусь к самому (самой) себе	,744	,306	-,094
В целом я удовлетворен (удовлетворена) собой	,742	,331	-,110
Иногда я чувствую себя бесполезным (бесполезной)	,797	-,239	-,229
Я хотел (хотела) бы относиться к себе с большим уважением	,768	-,016	-,317
Иногда мне кажется, что я нехороший человек	,803	-,066	-,202

Метод выделения: Анализ методом главных компонент.

а. Извлеченных компонент: 3

Однако работать с такой таблицей по-прежнему не очень удобно. Как видим, нужно долго выискивать в ней самые большие корреляционные значения, для того чтобы понять, к какому фактору больше относится та или иная переменная. То есть результат разбивки всех переменных на три фактора не столь очевиден.

Чтобы результат построения факторной модели был более нагляден, сделаем следующие действия. Во-первых, уберем из таблицы значения коэффициентов корреляции меньше 0,5 (в каждом случае при определении порога значимости исследователь должен исходить из логики исследования и принципа целесообразности). Их можно не учитывать, так как они слишком малы (об этом подробно написано в Лекции 5). Во-вторых, воспользуемся методом вращения факторов и увидим наглядно его результативность.

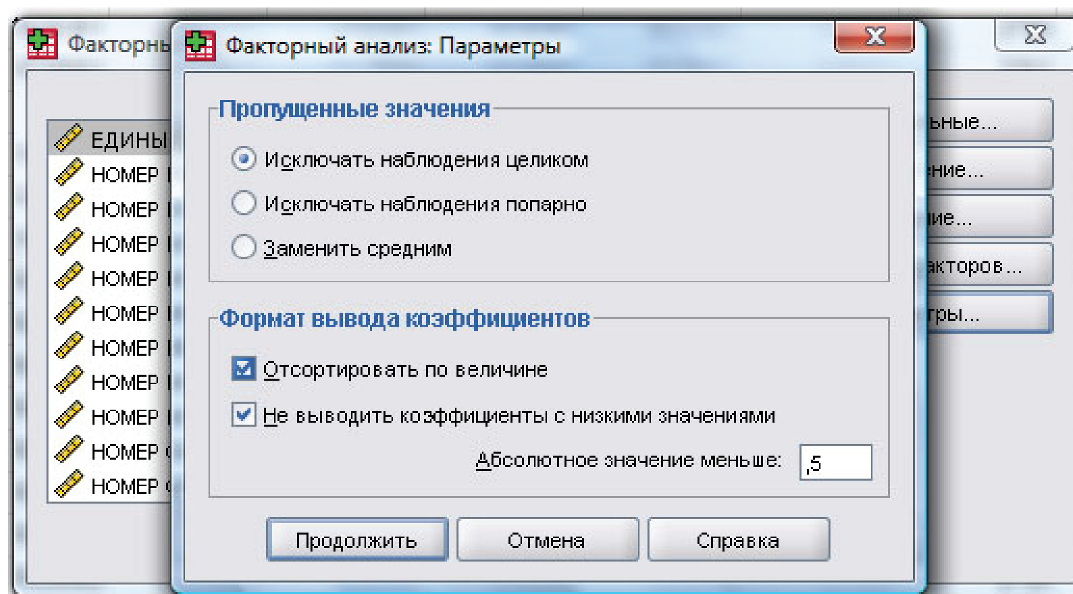


Рис. 9.7. Диалоговое окно «Факторный анализ: Параметры»

Для того чтобы из таблицы убрать слишком маленькие значения факторных нагрузок обратимся к функциональной кнопке «Параметры…» диалогового окна «Факторный анализ» (рис. 9.2). В появившемся новом диалоговом окне «Факторный анализ: Параметры» отметим в разделе «Формат вывода коэффициентов» пункты – «Отсортировать по величине» и «Не выводить коэффициенты с низкими значениями меньше 0,5» и нажмем кнопку «Продолжить».

Для того чтобы воспользоваться вращением факторов, выберем кнопку «Вращение…». На экране появится диалоговое окно следующего вида (см. рис. 9.8).

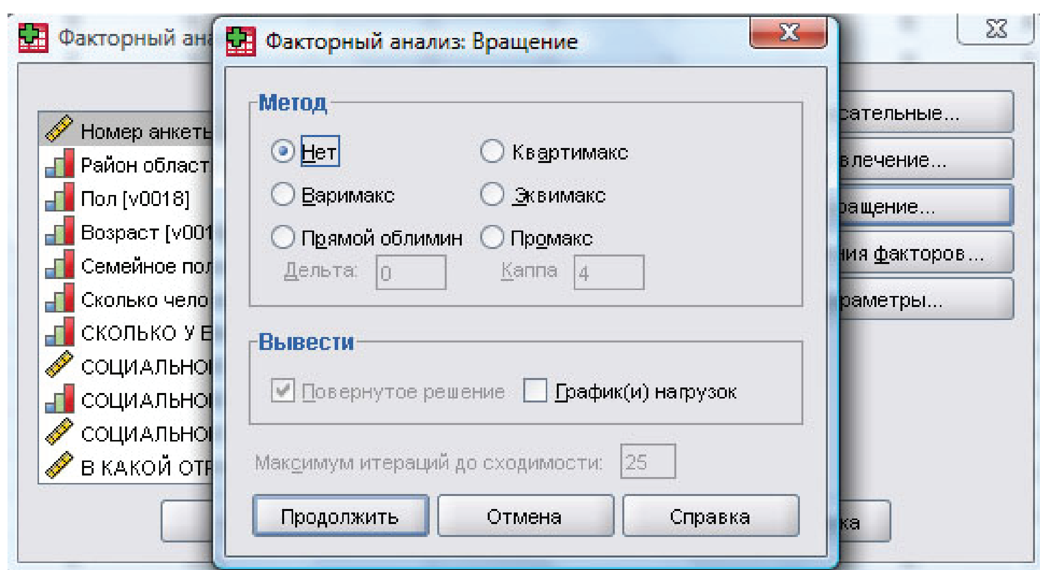


Рис. 9.8. Диалоговое окно «Факторный анализ: Вращение»

Здесь можно выбрать один из имеющихся способов вращения факторов. Как правило, вращение нужно для того, чтобы упростить интерпретацию факторов или переменных. В SPSS исследователю предложено несколько вариантов вращения факторов – варимакс, облимин, квартимакс, эквимакс, промакс.

Варимакс – метод ортогонального вращения факторов, минимизирующий число переменных с высокими нагрузками на каждый фактор, что существенно упрощает интерпретацию факторов.

Прямой облимин – метод косоугольного (неортогонального) вращения факторов.

Квартимакс – метод вращения факторов, минимизирующий число факторов, требуемых для объяснения каждой переменной, что существенно упрощает интерпретацию наблюденных переменных.

Эквимакс – метод вращения факторов, объединяющий методы варимакс, упрощающий факторы, и квартимакс, упрощающий переменные. При

его использовании минимизируется число переменных с большими факторными нагрузками и число факторов, требуемых для объяснения каждой переменной.

Промакс – метод косоугольного вращения, применяемый для факторной модели, в которой факторы могут коррелировать между собой.

Выбор конкретного способа вращения всегда остается за исследователем, который исходит из задач своего исследования, то есть из желаемых результатов.

Для нашего примера мы выберем варимакс вращение для того, чтобы максимально упростить интерпретацию выделенных факторов. Теперь вместо одной матрицы факторных нагрузок SPSS выводит две – простую (табл. 9.6) и повернутую (табл. 9.7). Напомним, что в таблице не выводятся значения меньше 0,5. Сравнение этих двух таблиц наглядно показывает результативность вращения факторов с точки зрения выяснения их содержательного наполнения и интерпретации. В таблице 9.7 уже отчетливо видно, какие именно переменные попадают в каждый из факторов.

Таблица 9.6

Матрица компонент[a]

	Компонента		
	1	2	3
Я не могу справиться со своими проблемами	,788		
Иногда я чувствую, что мной помыкают в жизни	,779		
Я мало могу влиять на то, что со мной происходит	,736		
Я всегда могу выполнить задуманное	,666		
Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни	,782		
То, что со мной произойдет в будущем, во многом зависит от меня	,703		
То, что я могу сделать, мало что изменит в моей жизни	,748		
Я думаю, что я ничем не хуже других	,773		
Я считаю, что у меня есть много хороших качеств	,787		
В общем, мне кажется, что я неудачник (неудачница)	,787		
Я могу все делать не хуже других	,766		
Я думаю, что мне особенно нечем гордиться	,778		
Я хорошо отношусь к самому (самой) себе	,744		
В целом, я удовлетворен (удовлетворена) собой	,742		
Иногда я чувствую себя бесполезным (бесполезной)	,797		
Я хотел (хотела) бы относиться к себе с большим уважением	,768		
Иногда мне кажется, что я нехороший человек	,803		

Метод выделения: Анализ методом главных компонент.

а. Извлеченных компонент: 3

Таблица 9.7

Матрица повернутых компонента

	Компонента		
	1	2	3
Я не могу справиться со своими проблемами	,743		
Иногда я чувствую, что мной помыкают в жизни	,739		
Я мало могу влиять на то, что со мной происходит	,724		
Я всегда могу выполнить задуманное	,716		
Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни	,710		
То, что со мной произойдет в будущем, во многом зависит от меня	,696		
То, что я могу сделать, мало что изменит в моей жизни	,681		
Я думаю, что я ничем не хуже других		,779	
Я считаю, что у меня есть много хороших качеств		,751	
В общем, мне кажется, что я неудачник (неудачница)		,750	
Я могу все делать не хуже других		,679	
Я думаю, что мне особенно нечем гордиться		,648	
Я хорошо отношусь к самому (самой) себе			,667
В целом я удовлетворен (удовлетворена) собой			,659
Иногда я чувствую себя бесполезным (бесполезной)			,651
Я хотел (хотела) бы относиться к себе с большим уважением		,516	,596
Иногда мне кажется, что я нехороший человек			,593

Метод выделения: Анализ методом главных компонент.
Метод вращения: Варимакс с нормализацией Кайзера.

a. Вращение сошлось за 7 итераций.

Таким образом, на основе выяснения степени соответствия всех пар суждений были определены следующие три фактора:

Фактор 1 (Ф-1): суждения 1–7 (всего семь суждений-мнений);

Фактор 2 (Ф-2): суждения 8–12 (всего пять суждений-мнений);

Фактор 3 (Ф-3): суждения 13–17 (всего пять суждений-мнений).

Для того чтобы проанализировать типы, которые могут быть выделены на основе этих трех факторов, сначала необходимо проанализировать сами факторы.

Фактор 1 составили следующие суждения-мнения:

Я не могу справиться со своими проблемами (С-1);

Иногда я чувствую, что мной помыкают в жизни (С-2);

Я мало могу влиять на то, что со мной происходит (С-3);

Я всегда могу выполнить задуманное (С-4);

Я часто чувствую себя беспомощным перед проблемами, возникающими в моей жизни (С-5);

То, что со мной произойдет в будущем, во многом зависит от меня (С-6);

То, что я могу сделать, мало что изменит в моей жизни (С-7).

Согласие с четырьмя из этих суждений и несогласие с двумя из них (С-4 и С-6) в целом характеризует позицию социального бессилия, когда человек пасует перед жизненными трудностями, предпочитая двигаться по течению «реки жизни». То есть это позиция социальной пассивности и неверия в способность что-то изменить в своей жизни, вера в предопределенность, а фактор фиксирует *степень активности жизненной позиции*. Таким образом, Ф-1 имеет шкалу: социальная пассивность (социальное бессилие) или провиденциализм – социальная активность (инициативность).

В фактор 2 вошли суждения:

Я думаю, что я ничем не хуже других (С-8);

Я считаю, что у меня есть много хороших качеств (С-9);

В общем, мне кажется, что я неудачник (неудачница) (С-10);

Я могу все делать не хуже других (С-11);

Я думаю, что мне особенно нечем гордиться (С-12).

Суждения-мнения, составляющие данный фактор, характеризуют в целом позицию, направленную на самоуважение и самоудовлетворение при согласии с 8-, 9- и 11-м суждениями и несогласии с 10- и 12-м. Таким образом, по данному фактору можно зафиксировать *уровень самоудовлетворения респондентов*. Шкала Ф-2 содержит позиции от «полностью неудовлетворен собой» до «полностью удовлетворен собой».

Фактор 3 составили оставшиеся пять суждений, которые в целом отражают *характер самоуважения респондентов*, основанный на общественной полезности:

Я хорошо отношусь к самому (самой) себе (С-13);

В целом я удовлетворен (удовлетворена) собой (С-14);

Иногда я чувствую себя бесполезным (бесполезной) (С-15);

Я хотел (хотела) бы относиться к себе с большим уважением (С-16);

Иногда мне кажется, что я нехороший человек (С-17).

Таким образом, Ф-2 и Ф-3 очень близки по содержанию наполняющих их суждений-мнений. Только Ф-3 имеет обратную шкалу, то есть «+» должен быть присвоен отрицательным значениям, а «–» – значениям положительным. В этом случае объяснительная модель будет более наглядной.

Взаимосвязь между факторами видна и математически из матрицы преобразования факторов (таблица 9.8), которая выводится автоматически при использовании преобразования вращения. В ней показаны коэффициенты корреляции между самими факторами.

Для продолжения анализа исследуемой совокупности респондентов необходимо сохранить результаты факторного анализа в виде новых переменных. Воспользуемся кнопкой «Значения факторов…» в диалоговом окне

«Факторный анализ» и получим новое окно, которое представлено на рисунке 9.9. Поставим в нем галочку напротив функции сохранения значений как переменных и выберем метод сохранения – регрессию. После нажатия кнопок «Продолжить» и «ОК» в нашей базе данных появятся три новых переменных – fac1_1, fac2_1, fac3_1 – в каждой из которых содержится значение соответствующего фактора для каждого наблюдения (респондента).

Таблица 9.8

Матрица преобразования компонент

Компонента	1	2	3
1	,625	,591	,511
2	,761	-,607	-,229
3	,175	,532	-,829

Метод выделения: Анализ методом главных компонент.
Метод вращения: Варимакс с нормализацией Кайзера.

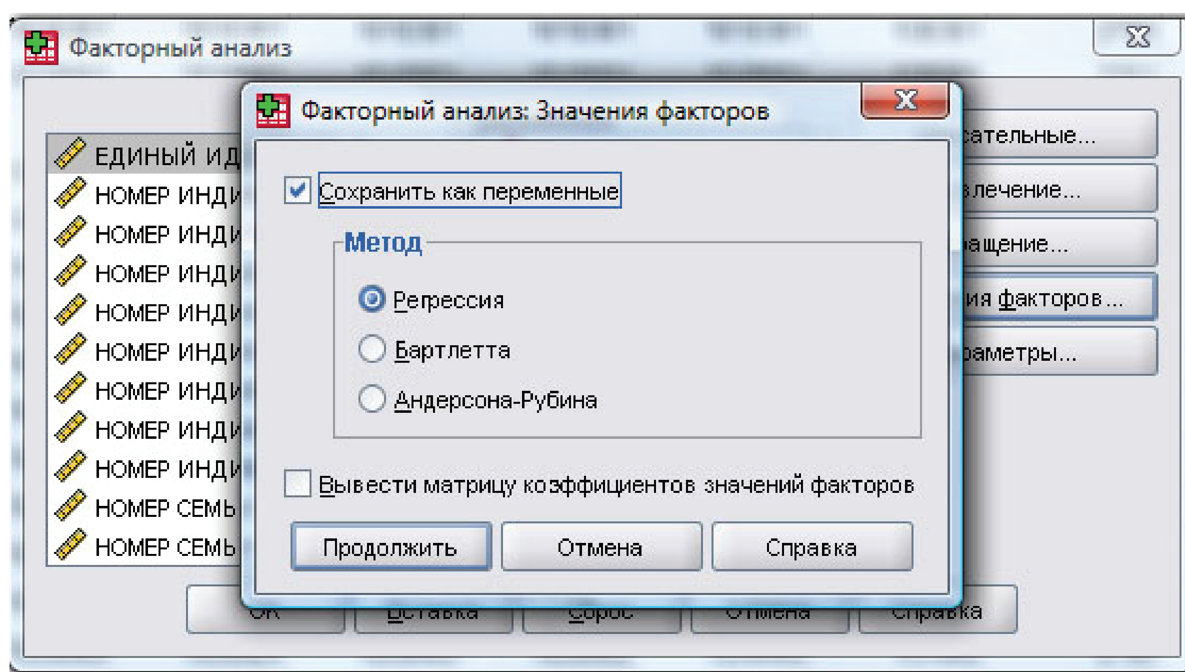


Рис. 9.9. Диалоговое окно «Факторный анализ: Значения факторов»

Используя в дальнейшем значения этих переменных, мы можем сгруппировать тем или иным образом всех опрошенных респондентов. На основе различных комбинаций этих трех факторов могут быть выделены восемь типов социально-психологической самоидентификации респондентов, которые представлены в таблице 9.9. В этой таблице приводится общее значение факторов в их сочетании, формирующем тот или иной тип самоидентификации респондента.

Таблица 9.9

Соотношение факторов Ф-1, Ф-2 и Ф-3 в различных типах социально-психологической самоидентификации респондентов

	Ф-1 *степень активности жизненной позиции*	Ф-2 *уровень самоудовлетворения респондентов*	Ф-3 *характер самоуважения респондентов*
1-й тип	+	+	+
2-й тип	–	+	+
3-й тип	–	–	+
4-й тип	–	–	–
5-й тип	+	–	–
6-й тип	+	+	–
7-й тип	–	+	–
8-й тип	+	–	+

1-й тип можно охарактеризовать как тип человека с активной жизненной позицией, считающего, что именно от него самого как никого другого зависят его жизненные успехи и неудачи. При этом люди, относящиеся к этому типу, выражают высокую степень самоудовлетворения, считают себя не хуже других, никогда не сомневаются в том, что они хорошие люди и могут быть полезны обществу и окружающим. Этот тип может быть условно назван *активисты*, в нашей выборке их 12,0% (см. рис. 9.10).

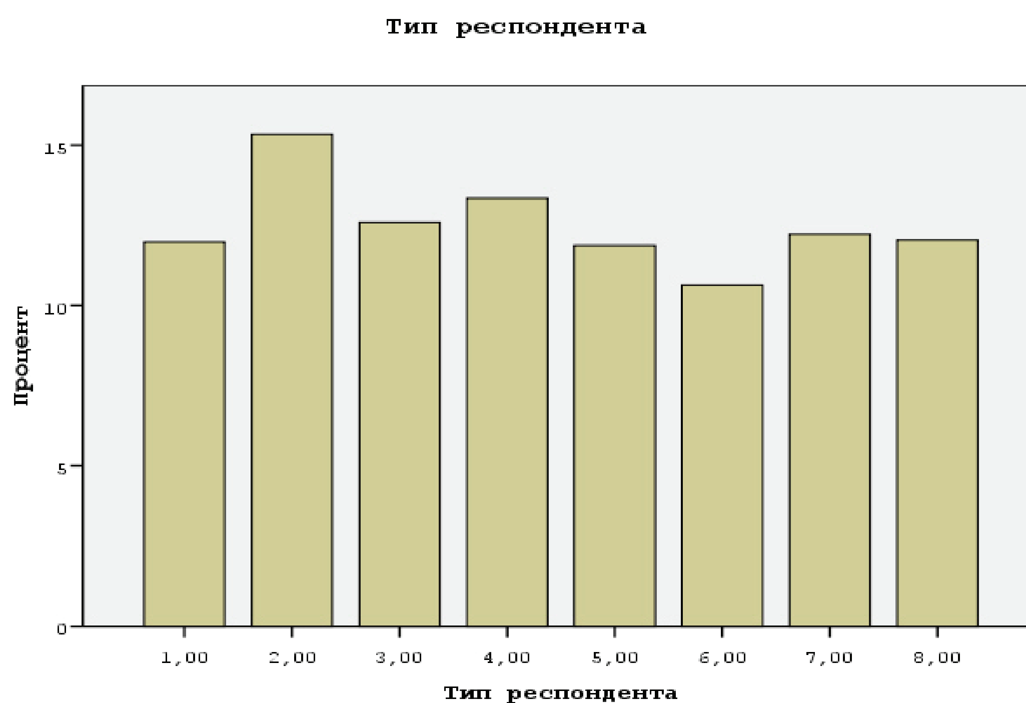


Рис. 9.10. Типы социальной самоидентификации респондентов на основе трехфакторной модели

2-й тип в отличие от 1-го не демонстрирует сколько-нибудь активной жизненной позиции, не склонен считать себя кузнецом собственной судьбы, однако это не мешает ему быть в ладу с самим собой, уважать себя и чувствовать свою пользу для общества. Такой тип респондентов мы условно назовем *фаталистами*, их несколько больше – 15,3%, и они являются самой многочисленной группой.

3-й тип еще более сложен. Он уже не верит в возможность изменить к лучшему свою жизнь, у него низкая самооценка и вера в собственные силы, но при этом вовсе не обязательно, что он считает себя неудачником или думает, что ему нечем гордиться. Такой тип мы назовем *пофигисты* (12,6%).

4-й тип характеризуется достаточно низким уровнем самооценки, отсутствием веры в собственные силы и возможность повлиять на свое положение в обществе и улучшить свои позиции. Кроме того, для данного типа характерно ощущение собственной бесполезности и даже беспомощности. Такой тип может быть назван полным *пессимистом*, доля коих в выборке составляет 13,3%.

5-й тип представляет другую комбинацию рассматриваемых факторов, при которой индивид достаточно низко оценивает себя, демонстрируя высокую степень неудовлетворенности собственными успехами, и зачастую считает себя нехорошим (или недостаточно хорошим) человеком, но при этом не теряет лицо и считает, что ему вполне по силам переломить ситуацию. Такой тип представляет собой *оптимиста*, к которым относится каждый восьмой респондент (11,9%).

6-й тип, выделенный нами на основе процедуры факторного анализа, характеризуется достаточно активной жизненной позицией и высоким уровнем самоудовлетворения. При этом такой человек демонстрирует низкую степень самоуважения, что на первый взгляд кажется несколько противоречивым, однако не является таковым для каждого десятого (10,6%) респондента. Такое возможно, когда человек чувствует, что он активно меняет свою судьбу, считает себя не хуже других, но считает, что его успехи на этом пути пока недостаточны, то есть является *прагматиком.*

Для *7-го типа* характерна вера в невозможность существенно повлиять на собственную судьбу, ощущение беспомощности и отсутствие поводов для гордости собой при одновременном спокойном отношении к этому, с учетом субъективного ощущения, что и у окружающих дела в этом отношении не лучше. Таким образом, 12,2% респондентов, относящихся к данному виду, являют собой пример социальных *конформистов*, живущих по принципу «мне хорошо, если остальным не лучше».

8-й тип отличается от всех остальных тем, что при достаточно активной жизненной позиции и высокой степени самоуважения они не демонстрируют высокой степени удовлетворенности своими успехами, считая,

что они скорее менее удачливы, чем многие из окружающих. Такой тип респондентов, включающий 12,0% наблюдений, мы условно назовем *нонконформистами*, не желающими мириться с такой ситуацией.

Таким образом, выделенные на основе трех факторов восемь типов самоидентификации респондентов позволяют получить интересные данные о структуре совокупности и открывают дополнительные возможности интерпретации данных.

На приведенном примере мы показали возможности использования факторного анализа в практике социологических исследований. Исследователю следует помнить, что сам по себе факторный анализ бессмыслен, если его результаты нельзя логично проинтерпретировать. Поэтому при проведении факторного анализа исследователь должен взвешивать все «за» и «против» на каждом этапе осуществления данного вида анализа – от отбора переменных и определения числа факторов до выбора метода анализа и способов вращения (в случае необходимости) матриц факторных значений. Только последовательное и продуманное осуществление всех этапов данного вида анализа может дать содержательный результат.

Вопросы и задания

1. Какова основная идея факторного анализа?
2. Какие методы факторного анализа можно использовать в SPSS? В чем разница между ними?
3. Как установить оптимальное число факторов?
4. Для чего в SPSS применяется вращение факторов? Какие процедуры вращения можно использовать для социологических данных?
5. Какова специфика интерпретации результатов факторного анализа?

Список литературы

1. Бессокирная, Г. П. Факторный анализ: традиции использования и новые возможности / Г. П. Бессокирная // Социология: 4М. – 2000. – № 12. – С. 142–153.

2. Благуш, П. Факторный анализ с обобщениями / П. Благуш. – М.: Финансы и статистика, 1989. – 248 с.

3. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цёфель. – СПб.: ДиаСофтЮП, 2005. – 608 с.

4. Гибсон У. Факторный, латентно-структурный и латентно-профильный анализ / У. Гибсон // Математические методы в социальных науках. – М.: Прогресс, 1973. – С. 9–41.

5. Крыштановский, А. О. Анализ социологических данных / А. О. Крыштановский. — М.: Изд-во «ГУ ВШЭ», 2007. — 281 с.

6. Окунь, Я. Факторный анализ / Я. Окунь. – М.: Статистика, 1974. – 200 с.

7. Попов, А. В. Основы проведения факторного анализа социально-экономического развития региона с использованием программного комплекса SPSS (на примере Алтайского края) // Вестник Российской экономической академии им. Г.В. Плеханова. – 2010. – №5. – С. 81–88.

8. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.

Лекция 10. Обзор основных статистических пакетов

В условиях информационного общества, когда в социальной действительности циркулирует много разнообразной информации, проводится огромное количество социологических, маркетинговых и других исследований, имеется насущная необходимость в тщательном анализе данных, базирующемся на методах математической статистики.

Большая роль в решении этой проблемы отводится современным информационным технологиям интеллектуального и статистического анализа данных. Выявление основных тенденций общественного мнения по социально значимым проблемам, построение социальных портретов различных слоев общества, оценка кредитных и страховых рисков, прогнозирование тенденций на финансовых рынках, оценка объектов недвижимости, построение профилей потенциальных покупателей определенного товара, анализ продуктовой корзины вот далеко не полный перечень задач, успешно решаемых с помощью систем интеллектуального и статистического анализа данных.

Системы интеллектуального анализа предназначены для автоматизированного поиска ранее неизвестных закономерностей в имеющихся в распоряжении социолога или управленца данных с последующим использованием полученной информации для подготовки решений. Помимо статистических методов, базовыми инструментами анализа в таких системах являются нейронные сети, деревья решений и индукция правил. Однако несмотря на то, что в последние годы рынок программных продуктов этого типа активно развивается, они все еще недоступны по цене предприятиям среднего и малого бизнеса. В то же время компаниям такого размера, как правило, не требуется столь мощный аналитический инструментарий, предлагаемый этими системами.

Более доступными средствами анализа данных на сегодняшний день являются статистические программные продукты (СПП). В мировой практике компьютерные системы статистического анализа и обработки данных широко применяются как в исследовательской работе в области экономики, так и в практической деятельности аналитических, маркетинговых и плановых отделов банков, страховых компаний, производственных и торговых фирм. В последние годы заметно возрос спрос на СПП и в нашей стране.

СПП позволяют решить широкий спектр задач «разведочного» анализа данных, статистического исследования зависимостей, планирования экспериментов, анализа временных рядов, анализа данных нечисловой природы и т.д. Настоящие методические разработки посвящены вопросам корреля-

ционно-регрессионного анализа статистических связей с использованием одного из самых популярных в России статистических программных продуктов – пакета STATISTICA, функционирующего в среде Windows.

Рынок СПП необычайно разнообразен. Существует около тысячи распространяемых на мировом рынке пакетов, решающих задачи статистического анализа данных в среде DOS, OS/2 или Windows.

Как ориентироваться в этом многообразии, если даже справочники, содержащие только краткие описания пакетов, представляют из себя объемные тома? Большую часть статистических пакетов можно разбить на две группы – это статистические пакеты общего назначения и специализированные программные продукты.

- Универсальные пакеты (или пакеты общего назначения) предлагают широкий диапазон статистических методов. В них отсутствует ориентация на конкретную предметную область. Они обладают дружественным интерфейсом. Из зарубежных универсальных пакетов наиболее распространены BAS, SPSS, Systat, Minilab, Statgraphics, STATISTICA.
- Специализированные пакеты, как правило, реализуют несколько статистических методов. Также они могут использовать методы, применяемые в конкретной предметной области. Чаще всего это системы, ориентированные на анализ временных рядов, корреляционно-регресионный, факторный или кластерный анализ. Применять такие пакеты целесообразно в тех случаях, когда требуется систематически решать задачи из этой области, для которой предназначен специализированный пакет, а возможностей пакетов общего назначения недостаточно. Из российских пакетов более известны STADIA, Олимп, Класс-Мастер, КВАЗАР, Статистик-Консультант; американские пакеты – ODA, WinSTAT, Statit и т.д.

Требования к СПП

Статистический пакет в идеале должен удовлетворять таким требованиям, как:

- модульность;
- ассистирование при выборе способа обработки данных;
- использование простого проблемно-ориентированного языка для формулировки задания пользователя;
- автоматическая организация процесса обработки данных и связей с модулями пакета;
- ведение банка данных пользователя и составление отчета о результатах проделанного анализа;

• диалоговый режим работы пользователя с пакетом;

• совместимость с другим программным обеспечением.

Следует заметить, что СПП постоянно развивается, постоянно создаются все новые варианты пакета, все в большей степени удовлетворяющие перечисленным выше требованиям. При этом на каждом этапе развития пакет, с одной стороны, должен представлять собой готовую к использованию программную продукцию, а с другой – входить составной частью в более поздние стадии развития пакета.

В состав методоориентированных СПП могут входить следующие функциональные блоки:

I. Блок описательной статистики и разведочного анализа исходных данных предусматривает:

- анализ смешанной природы многомерного признака и унификацию записи исходных данных;
- анализ резко выделяющихся наблюдений;
- восстановление пропущенных наблюдений;
- проверку статистической независимости наблюдений;
- определение основных числовых характеристик и частотную обработку исходных данных (построение гистограмм, полигонов частот, вычисление выборочных средних, дисперсий);
- статистическое оценивание параметров;
- вычисление модельных законов распределения вероятностей (нормального, биномиального, Пуассона, хи-квадрат и др.);
- визуализацию анализируемых многомерных статистических данных и др.

II. Блок статистического исследования зависимостей предполагает:

- корреляционно-регрессионный анализ;
- дисперсионный и ковариационный анализ;
- планирование регрессионных экспериментов и выборочных обследований;
- анализ временных рядов (предварительный анализ временных рядов; выявление тренда временного ряда; выявление скрытых периодичностей; спектральный анализ временного ряда; анализ случайных остатков временного ряда; проверка статистических гипотез: о стационарности ряда, независимости его членов, об адекватности «подгоняемой» модели) и др.

III. Блок классификации и снижения размерности включает:

- дискриминантный анализ;
- статистический анализ смесей распределений;

- кластер-анализ;
- снижение размерности в соответствии с критериями внешней информативности и автоинформативности и нек. др.

IV. Блок методов статистического анализа нечисловой информации и экспертных оценок. Среди используемого в этом блоке математико-статистического инструментария:

- анализ таблиц сопряженности;
- лог-линейные модели;
- субъективные вероятности;
- логит- и пробит-анализ;
- раговые методы и т.п.

V. Блок планирования эксперимента и выборочных обследований.

VI. Блок вспомогательных программ предусматривает статистическое моделирование на ЭВМ.

Рассмотрим более подробно наиболее распространенные универсальные статистические пакеты.

Система SAS

Информация об использовании системы SAS занимает одно из ведущих мест в публикациях, посвященных исследованию качества жизни. Алгоритмы шкалирования опросников зачастую распространяются в виде командного скрипта на встроенном языке системы.

Система SAS известна с 1976 г. и способна работать под управлением практически любой операционной системы (ОС). Установка SAS на компьютер приводит к инсталляции своей собственной операционной системы, которая, однако, способна обмениваться данными из приложений, работающих под управлением других ОС.

SAS включает свыше 20 различных программных продуктов, объединенных друг с другом *средствами доставки информации* (Information Delivery System, или IDS, так что весь пакет иногда обозначается как SAS/IDS). Благодаря IDS пользователю достаточно поставить на свой компьютер, кроме ОС, систему SAS, и этим можно ограничиться – все остальные функции типа задач, решаемых на основе Excel, Word, любой из СУБД и др. полностью возьмет на себя SAS/IDS. Традиционно сложилось, что основными отечественными пользователями системы являются предприятия ВПК, крупные бизнесмены (некоторые банки, включая Центробанк, биржи, торговые фирмы), некоторые атомные станции, крупнейшие медицинские и геофизические центры, крупные государственные структуры.

Основным достоинством SAS является непревзойденная мощность по набору статистических алгоритмов среди универсальных пакетов. Кроме того, SAS предоставляет пользователю возможность подключения собственных оригинальных алгоритмов.

Использованием SAS возможно решить практически любые задачи как систематизации данных, так и практически любого вида статистического анализа. Однако, высокая стоимость системы и малая распространенность ее в России делает ее малоизвестной среди отечественных специалистов, занимающихся исследованием качества жизни.

Универсальная статистическая система SYSTAT

Универсальная статистическая система SYSTAT разработана одноименной фирмой, которая в сентябре 1994 г. была поглощена корпорацией SPSS. Главное достоинство пакета – исключительно широкий диапазон и глубина проработки функционального наполнения. Здесь есть широкие возможности и для слабо подготовленного в статистике пользователя, и для достаточно искушенного статистика. Для исследователя качества жизни этот программный продукт представляет интерес благодаря наличию алгоритмов анализа шкал опросников, таких как анализ внутреннего постоянства, многомерное шкалирование, классический и логит-анализ пунктов шкалы.

Пакет MINITAB

Пакет MINITAB развивается более 20 лет и широко известен в США, где он является одним из основных учебных пакетов. Пакет работает как под ОС Windows, так и на компьютерах Macintosh. MINITAB хорошо продуман по разделу описательной (дескриптивной) статистики, хорошо сконструирован и управляется с помощью удобного меню, или, по желанию пользователя, через команды, составлять которые помогают диалоговые окна пакета. Часто используемые команды можно запускать по их первой букве. Общее число команд превышает 200. Можно составлять специальные макросы для выполнения последовательностей команд. Импорт/экспорт данных из других Windows-приложений делается через стандартный буфер обмена. В пакете имеются разнообразные возможности по управлению данными.

Пользователь MINITAB при исследовании качества жизни может легко и быстро решать практически все типовые задачи, в основном из области получения описательных статистик и сравнения групповых средних, анализа временных рядов. Если на этапе создания и валидации опросника исследования качества жизни требуется применение методов многомерной статистики, то MINITAB позволяет находить главные компоненты или даже про-

водить стандартный линейный или далее квадратичный дискриминантный анализ, использовать алгоритмы факторного и кластерного анализа.

Кроме того, MINITAB позволяет получать множество хороших и сложных полноцветных графиков. В плане характеристики мощности MINITAB достаточно силен и разнообразен, поэтому говорят, что первые четыре буквы пакета скорее надо поменять на Maxi.

Пакет Statistica 6.0

Пакет Statistica 6.0 не стоит использовать пользователю-новичку в статистике, так как он предполагает владение специальной терминологией. Тем не менее на отечественном рынке этот пакет пользуется популярностью благодаря высокой активности фирмы-разработчика Statsoft и дилера в России — Softline, способствующих популяризации пакета.

Ряд авторов считает, что пакет Statistica является хорошо сбалансированным по соотношению «мощность/удобство». Наличие широкого спектра функциональных алгоритмов делает его достаточно привлекательным для статистиков-профессионалов. В частности, он включает в себя ряд непараметрических методов анализа, методы многомерного анализа: дискриминантного, факторного кластерного, логлинейного и др. В области исследования качества жизни Statistica 6.0 предоставляет возможности анализа шкал и пунктов, а также обладает развитым блоком анализа мощности и необходимого количества наблюдений.

Средства манипулирования исходными данными в пакете Statistica хорошо развиты. Данные относительно легко отредактировать, можно создавать новые переменные («признаки»), выбирать отдельные наблюдения или «вырезать» подмножество данных по строкам и/или по столбцам таблицы «объект-признак». Благодаря обширной панели инструментов, для выполнения большинства манипуляций достаточно несколько щелчков мышки, так как почти для всех функций пакета здесь имеются пиктограммы.

Сильной стороной пакета является графика и средства редактирования графических материалов. В пакете представлены сотни типов графиков 2D или 3D, матрицы и пиктограммы. Предоставляется возможность разработки собственного дизайна графика.

Средства управления графиками позволяют работать одновременно с несколькими графиками, изменять размеры сложных объектов, добавлять художественную перспективу и ряд специальных эффектов, разбивку страниц и быструю перерисовку. Например, 3D-графики можно вращать, накладывать друг на друга, сжимать или увеличивать. Передовая анимационная

техника позволяет увидеть на графиках, какие точки изменились под влиянием изменений в одной из переменных.

Российский статистический пакет STADIA

Пакет STADIA разработан ведущими специалистами Московского государственного университета им. М. В. Ломоносова (главный разработчик — А. П. Кулаичев) совместно с НПО «Информатика и компьютеры». Первая версия пакета была создана в конце 70-х гг. для БЭСМ-6. С тех пор пакет постоянно модифицировался, пополняя свои функциональные и сервисные возможности.

Пакет STADIA является единственным российским статистическим пакетом, представленном на рынке, который можно отнести к классу универсальных пакетов, то есть в нем представлены все самые распространенные методы статистического анализа данных от описательной статистики и проверки различных гипотез до анализа временных рядов и контроля качества, а также многомерных (факторный, кластерный, дискриминантный анализ, шкалирование) и непараметрических методов анализа. Таким образом, пакет подходит для решения практически всех задач, встречающихся в исследовании качества жизни.

Пакет STADIA, в отличие от SAS и SPSS, не поддерживает обработку миллионов наблюдений, но прекрасно справляется с данными выборочных обследований нескольких сотен или тысяч респондентов. Пакет ориентирован на конкретные статистические расчеты и построение сопутствующих графиков во всех областях прикладной статистики, снабжая пользователя попутно всей необходимой информацией о работе статистических процедур.

В настоящее время пакет используется в учебном процессе и научно-практической работе более чем в 150 университетах России, включая 17 университетов медицинского профиля. Среди пользователей пакета — не только ведущие медицинские центры страны (НИИ им. Сербского, НИИ педиатрии РАМН, НИИ дефектологии, институт медико-биологических проблем, НИИ медицинского приборостроения и др.), но и поликлиники, больницы, медсанчасти городов: Москвы, Самары, Перми, Тулы, Уфы, Липецка, Архангельска, Кисловодска, Оренбурга, Бердянска и др.

Пакет STADIA простой в освоении, относительно недорогой и очень мощный инструмент статистического анализа данных ограниченных объемов. Он учитывает уровень подготовки российского пользователя, позволяет быстро найти необходимый метод обработки данных, представить результаты анализа в табличной и графической формах и продолжить их оформление в других средствах среды Windows (текстовых и графических редакторах).

STATGRAPHICS 5.1 for Windows

STATGRAPHICS включает более 250 статистических процедур, применяющихся в бизнесе, экономике, маркетинге, медицине, биологии, социологии, психологии, на производстве и в других областях. Каждой группе процедур соответствует собственное меню. Результаты представляются в табличной форме или на удобных для восприятия графиках.

Версия 5.1 обогащена диалоговой системой ввода данных из других приложений и выбора методов анализа. Уникальной особенностью STATGRAPHICS является процедура регрессионного анализа, где представлено сравнение полученной регрессионной зависимости с альтернативными моделями. При исследовании статистических связей между показателями качества жизни и клинико-лабораторными данными этот модуль может оказаться неоценимым.

Модуль Statistical Advisor, кратко поясняющий суть любого проведенного анализа, оказывает помощь в интерпретации результатов. Таким образом, STATGRAPHICS является полезным программным продуктов для исследования качества жизни, доступным как для начинающего исследователя, так и для совершенствующегося эксперта.

Пакет SPSS

Пакет SPSS предназначен в первую очередь для статистиков-профессионалов. Он включает развитый аппарат статистического анализа, соизмеримый по мощности с SAS. SPSS в настоящее время считают одним из лидеров среди универсальных статистических пакетов. Алгоритмы шкалирования опросников качества жизни распространяются также в виде скриптов на языке SPSS, причем научиться самостоятельно писать подобные алгоритмы способен даже специалист без начального программистского образования.

SPSS имеет удобные графические средства (более 50 типов диаграмм), а также развитые средства подготовки отчетов. Аналитические параметры отображаются на экране в виде простых и понятных меню и диалоговых окон. Новая контекстно-ориентированная справочная система содержит пошаговые инструкции для наиболее важных операций. В литературных источниках, посвященных исследованию качества жизни, упоминания об использовании SPSS встречаются практически наравне с упоминаниями о SAS.

Учитывая, что данный курс лекций посвящен именно статистическому пакету SPSS, остановимся более подробно на его истории и генезисе его возможностей для статистического анализа.

Студенты Норман Най (Norman Nie) и Дейл Вент (Dale Bent), специализировавшиеся в области политологии, в 1965 г. пытались отыскать в Стенфордском университете Сан-Франциско компьютерную программу, подходящую для анализа статистической информации. Вскоре они разочаровались в своих попытках, так как имеющиеся программы оказывались в большей или меньшей степени непригодными — неудачно построенными или не обеспечивали наглядность представления обработанной информации. К тому же принципы пользования менялись от программы к программе.

Они решили разработать собственную программу, со своей концепцией и единым синтаксисом. В их распоряжении тогда был язык программирования FORTRAN и вычислительная машина типа IBM 7090. Уже через год была разработана первая версия программы, которая еще через год, в 1967 г., могла работать на IBM 360. К этому времени к группе разработчиков присоединился Хэдлай Халл (Hadlai Hull).

Как известно из истории развития информатики, программы тогда представляли собой пакеты перфокарт. На это указывает и исходное название программы, которое авторы дали своему продукту: SPSS — это аббревиатура от Statistical Package for the Social Science.

В 1970 г. работа над программой была продолжена в Чикагском университете, а Норман Най основал соответствующую фирму — к тому моменту уже было произведено 60 инсталляций. Первое руководство для пользователей описывало 11 различных процедур.

Спустя пять лет SPSS была уже инсталлирована 600 раз, причем под разными операционными системами. С самого начала версиям программы присваивали соответствующие порядковые номера. В 1975 г. была разработана уже шестая версия (SPSS6). До 1981 г. были выпущены версии 7, 8 и 9.

Командный язык (синтаксис) SPSS в то время был еще не так хорошо развит, как сейчас, и, естественно, был ориентирован на перфокарты. Поэтому так называемые управляющие карты SPSS состояли из идентификационного поля (столбцы 1–15) и из поля параметров (столбцы 16–80).

В 1983 г. командный язык SPSS был полностью переработан, синтаксис стал значительно удобней. Чтобы отметить этот факт, программа была переименована в SPSSX, где буква X должна была служить как номером версии — началом нового исчисления, римскими цифрами, так и сокращением для «extended» (расширенный).

Так как применение перфокарт к этому моменту уже стало историей, то программа SPSS и информация, подлежащая обработке, сохранялись в отдельных файлах на винчестерах больших ЭВМ, которые тогда исполь-

зовались повсеместно. Год от года постоянно увеличивалось и количество процедур.

С появлением персональных компьютеров была разработана также и PC-версия SPSS, с 1983 г. появилась PC-версия SPSS\PC+. рассчитанная на MS-DOS. Позже, с момента основания в 1984 г. европейского торгового представительства в Горинхеме (Нидерланды), SPSS стал широко применяться и в Европе. В настоящее время это самое распространенное программное обеспечение для статистического анализа во всем мире.

Для того чтобы отразить возможность использования программы во всех областях, имеющих отношение к статистическому анализу, буква X вновь была удалена из названия марки, а исходной аббревиатуре присвоено новое значение: Superior Performance Software System (система программного обеспечения высшей производительности).

Если PC версия SPSS/PC+ была чуть усовершенствованной версией для больших ЭВМ, то SPSS для операционной системой Windows (SPSS for Windows) стала большим шагом вперед. Во-первых, эта версия SPSS обладает всеми возможностями версии для больших ЭВМ, во-вторых, за некоторыми немногочисленными исключениями, программой можно пользоваться, не имея особых знаний в области прикладного программирования. Вызов необходимых процедур статистического анализа происходит при помощи стандартной техники, применяемой в Windows, то есть с помощью мыши и соответствующих диалоговых окон.

Первая версия SPSS для Windows имела порядковый номер 5. Затем последовали версии 6.0 и 6.1 с некоторыми нововведениями в статистической и графической областях. Версия 6.1 была первой статистической программой для Windows, которая использовала 32-битную архитектуру Windows 3.1. Это можно было заметить по более высокой скорости выполнения вычислений. Усовершенствования коснулись также и интерфейса пользователя. В конце концов, была выпущена версия 6.1.3, которая уже могла работать и под Windows 95, и под NT.

В начале 1996 г. появилась 7-я версия SPSS, сначала как версия 7.0, а затем — как 7.5. Наряду с расширением возможностей в сфере статистики, разница между этими двумя версиями заключалась в том, что в версии 7.5 как меню, так и интерфейс программы были выполнены уже не только на английском, но и на других наиболее распространенных языках.

Самым весомым отличием версии 7 по отношению к предыдущим версиям является абсолютно новый подход к выводу информации на экран. Так, во-первых, получил новые очертания так называемый Viewer (Окно просмотра), во-вторых, более приятный внешний вид приобрели таблицы результа-

тов расчетов (мобильные таблицы). Появившаяся технология мобильных таблиц позволяет перестраивать полученные таблицы различным способами.

Если предшественница данной версии — версия 6.1.3 — могла работать как под старой Windows 3.1, так и под новой Windows 95 (NT), то SPSS версии 7 могла работать только при наличии Windows 95 (NT).

За версией 7.5 последовала версия 8.0, прогресс которой заключался в усовершенствовании графической оболочки. Возможность составления интерактивных графиков предоставляет ряд преимуществ по сравнению с традиционными графиками, которые являются стандартом для многих других пакетов.

Версия 9.0 включала в себя несколько новых статистических методов, в т. ч. многозначную логистическую регрессию, и несколько новых графических возможностей, расширяющих область интерактивных графиков.

У версии 10.0 SPSS несколько существенных отличий по сравнению с предыдущей версией — 9.0: было изменено строение Редактора данных и, благодаря закладкам Данные и Переменные, был облегчен переход между областями ввода данных и описания переменных. Таким образом, форма описания переменных была упрощена и соответствует теперь общепринятым стандартам, применяемым в сфере табличных расчетов. В области статистики был добавлен регрессионный анализ с категориальной целевой переменной.

В 11-ой версии SPSS основное внимание уделено расширению функциональных возможностей специальных модулей SPSS, таких как SPSS Categories, SPSS Advanced Models и другие.

Если раньше данная программа широко использовалась в таких «классических» областях науки и бизнеса, как биология, социология, психология, управление качеством производства, общие маркетинговые исследования и экономическое прогнозирование, то сейчас новую версию можно с успехом применять в таких актуальных специализированных областях, как маркетинг, основанный на использовании баз данных, Data Mining, Data Warehousing и другие. Особенного внимания заслуживает тот факт, что изменения, внесенные в модуль SPSS Regression Models, позволяют использовать SPSS при решении задач управления лояльностью клиентов (CRM). Отметим, что данная тема представляет собой один из наиболее популярных разделов современного практического маркетинга.

Отдельного упоминания заслуживает тот факт, что большинство наиболее популярных статистических методов прогнозирования, включенных в модуль SPSS Regression Models, позволяют работать с большим объемом недоступной информации. В математике в таком случае говорят о повышении

робастности метода, то есть его устойчивости по отношению к неопределенностям и существенным отклонениям от диапазона параметров, для которого разрабатывался метод. Такое повышение робастности весьма желательно в маркетинговых исследованиях и в социологии, где всегда присутствует большой объем отсутствующих или недостоверных данных. Небесполезно данное улучшение и в области управления качеством, где всегда существует компромисс между подробностью информации о производственном процессе и его усложнением.

Изменения коснулись и техники вычислений. Подобные изменения не сказываются на интерфейсе и прочих видимых функциональных особенностях программы, но, однако, они затрагивают вычислительное ядро, которое используется в ходе проведения конкретных расчетов. Здесь основное внимание было сосредоточено на повышении эффективности статистических алгоритмов, в некоторых случаях эффективность повысилась до 50 раз.

Эффективность одной из наиболее часто используемых статистических процедур, общей линейной модели (GLM), возросла в 10 раз, что, несомненно, скажется на обшей производительности при выполнении статистических исследований, особенно в области обработки больших массивов экспериментальных данных, которые возникают, например, в решении задач управления качеством, социологии и медицины.

В два раза выросла скорость выполнения самых массовых статистических процедур, таких как расчет дисперсии и вычисление средних. Можно смело сказать, что пользователь, который нуждается только в самых простых статистических методах, заметит именно двукратное повышение эффективности работы программы.

Повышение быстродействия особенно чувствуется в случае, когда речь идет о методах кластерного анализа, широко используемого в маркетинге, социологии, психологии и медицине, которые иногда требовали многочасовых расчетов даже на мощных компьютерах, для чего в предыдущих версиях SPSS был предусмотрен пакетный режим выполнения задач.

Следует отметить, что только одно столь существенное повышение производительности уже может быть основанием для выпуска новой версии программы. Снижение затрат времени, которое обеспечивает новая версия SPSS, позволяет более интенсивно использовать эту программу в практических маркетинговых исследованиях, анализировать большее количество вариантов, обрабатывать более широкие и представительные выборки. В результате издержки, связанные с исследованиями, падают, а степень достоверности информации повышается.

Изменения, которые были внесены в изобразительную и презентационную части программы, в основном затрагивают гибкость отображения результатов статистической обработки данных и включают несколько более показательных видов графиков. Например, при выводе информации о приближении данных с помощью выбранного метода аппроксимации, на графике приводится информация о том, насколько хорошо полученное приближение. Такая дополнительная возможность может оказаться весьма полезной для не слишком опытных пользователей или пользователей, не имеющих и не нуждающихся в глубокой математической подготовке. В целом изменения, которым подверглась графическая и презентационная часть программы, направлены на упрощение работы и облегчение интерпретации результатов вычислений неподготовленными пользователями.

Рассматривая изменения, внесенные в техническую часть программы, необходимо упомянуть, что новая версия SPSS способна конвертировать базовые и переносимые файлы программы SAS[21], своего наиболее мощного конкурента в области статистической обработки данных. Очень многие массивы общедоступной информации, имеющие отношения к маркетинговым исследованиям и социальной статистике, например данные по исследованию уровня жизни США и других стран (в том числе и России — знаменитый RLMS[22]), проводимые американскими исследователями, имеют формат переносимых файлов SAS.

Кроме того, следуя тенденции к превращению SPSS в мощное средство для проведения маркетинговых исследований и анализа разнородной информации, в 11-ой версии существенно возросло удобство доступа к различным форматам баз данных. В список поддерживаемых форматов теперь входят Sybase 11 и 12; Infomix 7.3+, 9.14: Infomix 2000 (9.20); UDB (DB2 6.1 и 7.1); SQL Server 2000; Oracle 8.06; Oracle! Releases 2 and 3 (8.1.6, 8.1.7). Улучшена связь с Microsoft Data Access pack. Более мощным стал язык запросов, появилась возможность на уровне запроса формировать и имена переменных и метки, что облегчает интерпретацию результатов и повышает их наглядность. Повысилась гибкость и функциональные возможности мобильных таблиц — это изменение затрагивает модуль SPSS Tables.

Начиная с 12-й версии, выходит русская версия программного обеспечения SPSS для Windows. Главными новшествами русской версии SPSS 12.0 являются:

- учебник на русском языке, позволяющий шаг за шагом освоить возможности SPSS и быстрее приступить к работе;

[21] См.: **www.sas.com**

[22] См.:**www.unc.edu**

- Репетитор по статистике на русском языке, помогающий в выборе нужной статистической или графической процедуры для конкретных данных и задач;
- Справка по SPSS Base и SPSS Tables на русском языке.

Среди основных новшеств работы с данными можно назвать следующие: новые возможности доступа к данным и управления данными (конструктор реструктуризации данных, новые инструменты «Свойства переменных» и «Копирование свойств данных», поиск дублирующихся наблюдений, визуальная категоризация, увеличение длины имени переменной с 8 до 64 знаков); новая процедура «Настраиваемые таблицы» в модуле SPSS Tables позволяет создавать таблицы и в процессе их построения видеть, как они будут выглядеть в конечном итоге.

В SPSS 13.0 были реализованы следующие новшества:

- Вывод результатов пополнился тремя новыми типами диаграмм: пирамиды населения, также называемыми зеркальными, или двойными, диаграммами; трехмерные гистограммы; точечные диаграммы, также называемые диаграммами плотности точек.
- Новые возможности отображения данных на диаграммах: Панели диаграмм (для большинства диаграмм, доступных в SPSS); Столбики ошибок для категориальных диаграмм позволяют включить в диаграммы информацию о достоверности данных; возможность сортировки категорий на диаграммах; гибкая работа с метками данных на диаграммах; диагональные опорные линии; усовершенствованный редактор диаграмм; новые возможности работы с шаблонами диаграмм
- Управление данными и выводом результатов: Конструктор даты и времени, существенно облегчающий расчеты и преобразования, в которых используются переменные дат и времени; увеличение максимальной длины текстовых переменных до 32767 байт; улучшена процедура автоматической перекодировки текстовых переменных в числовые переменные; сохранение агрегированных значений прямо в активном файле; возможность работы с Системой управления выводом (OMS) через интерфейс; чтение/запись файлов данных SAS 9; возможность беспрерывного выполнения синтаксиса, несмотря на возникающие ошибки; новая команда HOST для «выхода» в операционную систему и синхронного выполнения других приложений; усовершенствованные возможности экспорта результатов из SPSS в Microsoft PowerPoint

- Изменения, касающиеся анализа данных: новый дополнительный модуль SPSS Classification Trees, позволяющий непосредственно в SPSS для Windows строить деревья классификаций и решений, идентифицировать группы, находить взаимосвязи в данных и предсказывать будущие события. В дополнительный модуль SPSS Complex Samples теперь добавлены новые процедуры: Общая линейная модель (в которую, в частности, входит дисперсионный анализ) и Логистическая регрессия. В дополнительном модуле SPSS Tables появились возможности сортировки категорий по любой итожащей статистике в таблице, а также скрытия (и показа) категорий, входящих в подитоги. Возможности дополнительного модуля SPSS Categories существенно расширились за счет добавления процедуры Множественный анализ соответствий. В SPSS Regression Models появились новые методы для шагового отбора в Мультиномиальной логистической регрессии: метод скоринга и метод Вальда; а также новые критерии подгонки модели: информационный критерий Акаике (AIC) и Байесовский информационный критерий (BIC)

В SPSS 14.0 появилась возможность работать с новыми типами файлов (в том числе в формате Stata), открывать несколько наборов данных в рамках одной сессии, импортировать данные из SPSS Dimensions и т. д., создавать собственные диаграммы с помощью графического языка программирования (GPL) и нового инструмента Chart Builder. Новый модуль SPSS Data Validation позволяет автоматически проверять данные и получать более точные прогнозы, а усовершенствованный модуль SPSS Trends дает возможность автоматически выбрать модель, которая лучше всего подходит для прогнозирования данного временного ряда.

С помощью расширенных опций моделирования уравнений и приложения Amos 6.0 пользователи могут преобразовать сложные уравнения и связанные с ними модели в более простые.

Начиная с версии SPSS 16, файлы Вывода (Output) (*.spv) утратили совместимость с версиями, созданными в предыдущих версиях пакета (расширение spo). Это произошло в связи с переработкой форматов и способов вывода результатов работы SPSS for Windows, что вызвало необходимость создания нового типа файла. Чтобы обеспечить возможность просмотра файлов в формате spo в SPSS 16 for Windows и более поздних версиях, на CD с пакетом размещена бесплатная программа “Legacy Viewer”.

Помимо кроссплатформенности, в 16-й версии были представлены следующие новшества:

- новый интерфейс, более быстрый и интуитивно-понятный;
- расширенные аналитические возможности;
- усовершенствованные возможности программирования (в качестве скриптового языка выбран простой и мощный Python);
- усовершенствованные функции генерации отчетов;
- повышенное быстродействие на многоядерных процессорах.

Среди основных новшеств SPSS Statistics 17.0:

- возможность работать под Windows Vista (а также под Windows XP, Mac OS X и Linux);
- усовершенствованный Редактор синтаксиса;
- улучшенный Конструктор диаграмм и существенно расширенные возможности визуализации данных;
- процедура Оптимальной категоризации (в SPSS Data Preparation);
- нейронные сети (в SPSS Neural Networks);
- анализ RFM (в SPSS EZ RFM);
- процедуры Обобщенная линейная модель и Обобщенные уравнения (в SPSS Advanced Statistics);
- процедура анализа методом ближайшего соседа;
- поддержка языка скриптов Python, а также другие важные новшества и усовершенствования.

Между 2009 и 2010 гг. название программного обеспечения SPSS было изменено на PASW (Predictive Analytics SoftWare) Statistics. Это произошло после того, как 28 июля 2009 г. компания объявила, что она была приобретена компанией IBM за 1,2 млрд долл. США. По состоянию на январь 2010 г. компания стала называться «SPSS: An IBM Company». Таким образом, все последующие версии программы выходили уже под новым именем (в частности, 18-я и 19-я версии).

Вопросы и задания

1. Перечислите основные статистические пакеты, используемые для анализа данных социологических исследований, кратко охарактеризуйте их.
2. Кратко опишите генезис версий пакета SPSS.
3. Как вы считаете, почему практически во всех архивах социологических данных результаты исследований хранятся в формате программы SPSS?

Список литературы

1. Киштович А. Краткий обзор некоторых статистических подходов [Электронный ресурс] // Межнациональный центр исследования качества жизни. URL: http://www.quality-life.ru/metodologiya01.php#01

2. Официальный сайт компании SPSS [Электронный ресурс]. URL: http://spss.ru/.

Заключение

Умение обрабатывать и анализировать данные, полученные в результате социологических исследований, является необычайно важным для студента-социолога. Надеемся, что читатель не только получил общее представление о пакете SPSS и его возможностях, но и научился правильно использовать эти почти безграничные возможности в собственных интересах, и информация, представленная в этом курсе лекций, оказалась полезной с практической точки зрения.

Между тем сама программа SPSS совершенствуется, выходят новые ее версии, в пакет включаются новые методы анализа. Интересно, что когда работа над курсом лекций началась, последней была 17-я версия программы, а когда закончилась – в свет вышли уже 18-я и 19-я. В этих условиях ни один курс лекций и ни одно учебное пособие не может являться исчерпывающим в своем роде. Исследователям, молодым и уже зрелым, самостоятельно приходится прокладывать путь к знаниям и искать ответы на возникающие вопросы.

Вполне естественно, что и учебная литература со временем обновляется. Со своей стороны хочется отметить, что работа авторов в этом направлении продолжается. Надеемся, что со временем круг рассмотренных нами тем (наиболее важных для социологов) расширится, а сам материал станет более доступным с точки зрения восприятия и усвоения. В связи с этим хочется обратиться к читателю с просьбой о помощи в совершенствовании издания. Мы будем признательны вам за ваше мнение о содержании лекций, советы пожелания по их доработке. Просим присылать их по адресу: fshamil@mail.ru.

Учебное издание

Фарахутдинов Шамиль Фаритович
Бушуев Алексей Сергеевич

ОБРАБОТКА И АНАЛИЗ ДАННЫХ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ В ПАКЕТЕ SPSS 17.0
КУРС ЛЕКЦИЙ

Редактор *В. Н. Ионина*
Компьютерная верстка *М. В. Юркин*

Подписано в печать 21.09.2011. Формат 60x90 1/16. Усл. печ. л. 13,75.
Тираж 500 экз. Заказ № 330.

Библиотечно-издательский комплекс
федерального государственного бюджетного образовательного
учреждения высшего профессионального образования
«Тюменский государственный нефтегазовый университет».
625000, Тюмень, ул. Володарского, 38.

Типография библиотечно-издательского комплекса.
625039, Тюмень, ул. Киевская, 52.